

Le origini del Calcolo delle Probabilità: dalle sale da gioco alla formalizzazione assiomatica

La teoria delle probabilità rappresenta uno dei pilastri fondamentali della matematica moderna, eppure le sue origini sono curiosamente legate a questioni apparentemente frivole: i giochi d'azzardo, le scommesse e i problemi di equa divisione delle poste. Questa disciplina, che oggi permea ogni aspetto della scienza moderna dalla fisica quantistica all'intelligenza artificiale, ha iniziato il suo cammino nelle sale da gioco del XVII secolo, quando alcuni dei più grandi matematici dell'epoca si confrontarono con domande apparentemente semplici ma profondamente ricche di implicazioni teoriche.

Il percorso che ci ha portato dalla nozione intuitiva di "caso" alla rigorosa formalizzazione assiomatica di Andrej Kolmogorov è stato lungo e tortuoso, costellato di intuizioni brillanti, dibattiti filosofici e progressivi raffinamenti concettuali. Nel presente contributo esploreremo questa evoluzione in dettaglio, analizzando non solo i risultati matematici ma anche il contesto storico e culturale che li ha generati.

Capitolo I: Le Radici Antiche - Prima di una Teoria Formale

Sebbene si tenda a identificare la nascita del calcolo delle probabilità con la corrispondenza tra Pascal e Fermat nel 1654, sarebbe errato pensare che prima di quella data non esistesse alcuna riflessione sul caso e sull'incertezza. Gli esseri umani hanno sempre dovuto confrontarsi con l'incertezza, e tracce di ragionamenti probabilistici si trovano già nell'antichità.

I Romani utilizzavano gli astragali, piccoli ossi del tarso di alcuni animali, come dadi, e certamente avevano una comprensione empirica delle diverse probabilità associate ai loro lanci. Tuttavia mancava completamente una teoria matematica che potesse quantificare queste intuizioni. Aristotele stesso aveva riflettuto sulla natura del caso e della contingenza nella sua Fisica, distinguendo tra eventi necessari, impossibili e possibili, ma sempre in un contesto filosofico piuttosto che matematico.

Durante il Medioevo, alcuni studiosi arabi svilupparono importanti contributi nel campo della combinatoria, che sarebbe poi diventata uno strumento essenziale per il calcolo delle probabilità. Al-Khalil ibn Ahmad al-Farahidi (718-786) studiò sistematicamente le permutazioni delle parole arabe, mentre nel XIII secolo il matematico cinese Yang Hui esplorò il triangolo aritmetico che in Occidente sarebbe stato poi associato a Pascal.

Un passaggio cruciale si ebbe con Gerolamo Cardano (1501-1576), medico, matematico e giocatore accanito. Nel suo Liber de Ludo Aleae (Libro sui giochi d'azzardo), scritto intorno al 1564 ma pubblicato postumo solo nel 1663, Cardano

presentò alcuni dei primi tentativi di analisi matematica dei giochi di fortuna. Egli comprese che con un dado equo ciascuna delle sei facce ha uguale probabilità di apparire, e che quindi la probabilità di ottenere un particolare numero è $1/6$. Più significativamente, Cardano intuì che per eventi indipendenti le probabilità si moltiplicano: la probabilità di ottenere due sei consecutivi lanciando due dadi è $(1/6) \times (1/6) = 1/36$.

Tuttavia l'approccio di Cardano rimase largamente empirico e non si sviluppò in una teoria generale. Mancava ancora un linguaggio matematico adeguato e, soprattutto, mancava l'interesse della comunità matematica dell'epoca per questi problemi, considerati troppo legati al gioco e quindi poco rispettabili dal punto di vista intellettuale.

Capitolo II: Il 1654 - L'Annus Mirabilis della Probabilità

Il Problema del Cavaliere di Méré

La vera svolta nella storia del calcolo delle probabilità avvenne grazie a un aristocratico francese appassionato di gioco, Antoine Gombaud, Cavaliere di Méré. Nel 1654 il Cavaliere di Méré pose a Blaise Pascal due quesiti che si sarebbero rivelati di importanza fondamentale per lo sviluppo della teoria.

Il primo problema riguardava una questione che aveva tormentato de Méré nelle sue esperienze di gioco. Egli aveva notato empiricamente che scommettere di ottenere almeno un sei in quattro lanci di un dado sembrava essere vantaggioso, mentre scommettere di ottenere almeno un doppio sei in 24 lanci di due dadi sembrava essere svantaggioso. Questo lo lasciava profondamente perplesso, dato che il rapporto $24:4 = 6:1$ corrispondeva esattamente al rapporto tra le possibilità di ottenere un doppio sei (1 su 36) e quelle di ottenere un sei (1 su 6). La sua intuizione suggeriva che i due giochi dovessero essere equivalenti, ma l'esperienza al tavolo da gioco lo contraddiceva.

Pascal analizzò il problema con il rigore che sarebbe diventato caratteristico del nuovo calcolo. Per il primo caso, la probabilità di non ottenere un sei in un singolo lancio è $5/6$. Poiché i lanci sono indipendenti, la probabilità di non ottenere alcun sei in quattro lanci consecutivi è quindi $(5/6)^4 = 625/1296 \approx 0,4823$. Di conseguenza, la probabilità complementare di ottenere almeno un sei è $1 - (5/6)^4 = 671/1296 \approx 0,5177$, che è leggermente maggiore di $1/2$, rendendo effettivamente la scommessa favorevole al giocatore.

Per il secondo caso, la probabilità di non ottenere un doppio sei in un singolo lancio di due dadi è $35/36$. La probabilità di non ottenere alcun doppio sei in 24 lanci è quindi

$(35/36)^{24} \approx 0,5086$. Conseguentemente, la probabilità di ottenere almeno un doppio sei è $1 - (35/36)^{24} \approx 0,4914$, che è leggermente minore di $1/2$, rendendo la scommessa sfavorevole al giocatore.

Questa analisi dimostrò che l'intuizione del Cavaliere di Méré era corretta nel riconoscere una differenza tra i due giochi, ma per ragioni più sottili e matematicamente profonde di quanto egli avesse immaginato. Il rapporto semplice 6:1 non catturava la complessità del problema, che richiedeva invece la comprensione delle probabilità composte e del concetto di evento complementare. Questo esempio illustra magnificamente come l'intuizione pratica, pur potendo guidare verso la verità, necessiti del rigore matematico per essere pienamente compresa e giustificata.

Il Problema della Divisione delle Poste

Il secondo problema posto da de Méré a Pascal era ancora più profondo dal punto di vista teorico e aveva già una lunga storia di tentativi di soluzione. Si tratta del cosiddetto "problema della divisione delle poste" (*problème des partis*): supponiamo che due giocatori di uguale abilità stiano giocando una serie di partite, e che il primo a vincere n partite si aggiudichi l'intera posta. Se il gioco viene interrotto prematuramente quando un giocatore ha vinto a partite e l'altro b partite (con $a, b < n$), come dovrebbe essere divisa equamente la posta tra i due giocatori?

Questo problema era stato considerato già nel Medioevo da diversi matematici. Fra' Luca Pacioli, nel suo monumentale *Summa de Arithmetica, Geometria, Proportioni et Proportionalita* (1494), aveva proposto una soluzione basata sul semplice rapporto tra le partite già vinte: se un giocatore ha vinto a partite e l'altro b , la posta dovrebbe essere divisa nel rapporto $a:b$. Tuttavia questa soluzione è manifestamente insoddisfacente: non tiene minimamente conto del numero totale di partite necessarie per vincere il gioco. Per esempio, se il primo giocatore ha vinto 3 partite e il secondo 2, ma sono necessarie 100 vittorie per aggiudicarsi la posta, la divisione 3:2 ignora completamente il fatto che entrambi i giocatori sono ancora molto lontani dalla vittoria finale.

Cardano aveva proposto una soluzione leggermente migliore, considerando anche le partite che restano da giocare, ma anche la sua analisi era matematicamente incompleta e non forniva il metodo generale di soluzione. Il problema richiedeva un approccio completamente nuovo, che considerasse non solo la situazione attuale ma tutti i possibili sviluppi futuri del gioco.

Pascal iniziò una corrispondenza matematica con Pierre de Fermat nell'estate del 1654, e insieme i due matematici svilupparono due metodi diversi ma matematicamente equivalenti per risolvere il problema. La loro corrispondenza,

fortunatamente conservata e pubblicata postuma, rappresenta l'atto di nascita formale del calcolo delle probabilità come disciplina matematica.

L'Approccio di Pascal: Il Triangolo Aritmetico

Pascal affrontò il problema utilizzando quello che oggi chiamiamo "triangolo di Pascal", anche se la struttura combinatoria era già nota a matematici precedenti come Tartaglia e lo stesso Yang Hui. L'idea geniale di Pascal fu quella di considerare sistematicamente tutti i possibili esiti delle partite rimanenti e di contarli utilizzando le proprietà del suo triangolo aritmetico.

Supponiamo che manchino $r = n - a$ partite al primo giocatore e $s = n - b$ partite al secondo per raggiungere le n vittorie necessarie e aggiudicarsi la posta. Il numero totale di partite che dovrebbero essere giocate al massimo è $r + s - 1$, perché una volta che uno dei due giocatori raggiunge n vittorie totali, il gioco termina immediatamente. Tuttavia, per semplicità di calcolo, Pascal considerò tutte le $r + s - 1$ partite anche se nella realtà il gioco potrebbe terminare prima, osservando che questo non cambia il risultato finale poiché stiamo solo determinando quale giocatore raggiungerà per primo il traguardo.

In $r + s - 1$ partite future il numero totale di possibili sequenze di vittorie e sconfitte è $2^{(r+s-1)}$, assumendo che ogni partita sia equiprobabile. Di queste sequenze possibili, il primo giocatore vince il gioco complessivo se ottiene almeno r vittorie nelle $r + s - 1$ partite considerate. Dobbiamo quindi contare in quanti modi può ottenere esattamente k vittorie per ogni valore $k = r, r+1, \dots, r+s-1$.

Il numero di modi in cui possono verificarsi esattamente k vittorie in $r+s-1$ partite è dato dal coefficiente binomiale:

$$C(r+s-1, k) = (r+s-1)! / [k!(r+s-1-k)!]$$

La probabilità che il primo giocatore vinca il gioco complessivo è quindi:

$$P_1 = [\sum_{k=r}^{r+s-1} C(r+s-1, k)] / 2^{(r+s-1)}$$

Pascal sviluppò metodi eleganti per calcolare queste somme utilizzando le proprietà ricorsive del suo triangolo aritmetico, in particolare l'identità $C(n,k) = C(n-1,k-1) + C(n-1,k)$. La posta dovrebbe quindi essere divisa in modo che il primo giocatore riceva una frazione P_1 dell'ammontare totale, e il secondo giocatore riceva la frazione complementare $P_2 = 1 - P_1$.

L'Approccio di Fermat: L'Analisi Combinatoria

Fermat preferì un approccio leggermente diverso, basato sull'enumerazione esplicita e diretta di tutti i casi possibili. Per esempio, se al primo giocatore mancano 2 vittorie e al secondo ne mancano 3, Fermat considerò tutte le possibili sequenze di esattamente $2 + 3 - 1 = 4$ partite future.

Rappresentando con A una vittoria del primo giocatore e con B una vittoria del secondo, le $2^4 = 16$ sequenze possibili sono:

AAAA, AAAB, AABA, AABB, ABAA, ABAB, ABBA, ABBB,
BAAA, BAAB, BABA, BABB, BBAA, BBAB, BBBA, BBBB

Il primo giocatore vince il gioco complessivo se nella sequenza compaiono almeno 2 occorrenze della lettera A prima che compaiano 3 occorrenze della lettera B. Contando esplicitamente caso per caso, troviamo che il primo giocatore vince in 11 casi su 16, quindi la posta dovrebbe essere divisa equamente nel rapporto 11:5.

Sebbene questo metodo possa sembrare più elementare e meno sofisticato dal punto di vista matematico, ha il considerevole merito di essere estremamente chiaro, diretto e trasparente dal punto di vista concettuale. Fermat riconosceva esplicitamente che per casi più complessi questo approccio enumerativo diventa computazionalmente oneroso e praticamente irrealizzabile, ma lo considerava concettualmente più trasparente e quindi pedagogicamente superiore per comprendere la natura del problema.

Il Significato Profondo della Soluzione

Ciò che rende la soluzione di Pascal e Fermat così importante nella storia della matematica non è solo la correttezza matematica del risultato finale, ma il fatto che essa introduce implicitamente diversi concetti fondamentali che sarebbero diventati i pilastri della teoria delle probabilità:

1. Lo spazio degli eventi elementari: L'idea di considerare sistematicamente tutti gli esiti possibili di un esperimento aleatorio, rappresentati come le sequenze di vittorie e sconfitte.
2. Gli eventi equiprobabili: L'assunzione fondamentale che, in assenza di informazioni contrarie o di asimmetrie nel gioco, tutti gli esiti elementari hanno uguale probabilità di verificarsi.

3. La probabilità come rapporto: La definizione operativa di probabilità come rapporto tra il numero di casi favorevoli (quelli in cui si verifica l'evento di interesse) e il numero totale di casi possibili (assumendo equiprobabilità).

4. Il valore atteso: Implicitamente, la divisione della posta secondo le probabilità di vittoria introduce il concetto di valore atteso o speranza matematica, ossia la quantità che ci si aspetta di vincere "in media".

Questi concetti, che oggi ci sembrano ovvi e quasi banali, rappresentavano all'epoca una vera rivoluzione concettuale nel modo di pensare l'incertezza e di quantificarla matematicamente. Prima di Pascal e Fermat, si pensava all'incertezza in termini vaghi e qualitativi; dopo di loro, divenne possibile trattarla con il rigore e la precisione del calcolo matematico.

Capitolo III: Christiaan Huygens e il Primo Trattato Sistemático

La corrispondenza tra Pascal e Fermat circolò rapidamente negli ambienti matematici europei, suscitando grande interesse e curiosità. Nel 1657 il matematico e fisico olandese Christiaan Huygens venne a conoscenza dei problemi discussi dai due matematici francesi durante un viaggio a Parigi. Profondamente affascinato dalle questioni sollevate e dalla loro eleganza matematica, Huygens decise di svilupparle in modo sistematico e rigoroso, creando la prima esposizione organica della nascente teoria.

Nel 1657 pubblicò *De Ratiociniis in Ludo Aleae* (Sul Ragionamento nel Gioco dei Dadi), il primo trattato stampato interamente dedicato al calcolo delle probabilità. È importante notare che Huygens non aveva letto direttamente la corrispondenza Pascal-Fermat, ma aveva ricevuto solo una descrizione orale e sommaria dei problemi e delle loro soluzioni. Risolse quindi i problemi in modo completamente indipendente, un fatto notevole che testimonia sia la sua straordinaria abilità matematica sia la naturalità delle idee sottostanti una volta che il problema è stato formulato correttamente. Inoltre, Huygens aggiunse numerose considerazioni originali e nuovi problemi che estendevano significativamente la portata della teoria.

Il Concetto di Aspettativa

Il contributo più importante e duraturo di Huygens fu l'introduzione esplicita e formale del concetto di "aspettativa" (*expectatio*), che oggi chiamiamo valore atteso o speranza matematica. Huygens formulò il seguente principio fondamentale, che pose alla base di tutta la sua trattazione:

"Se io ho p possibilità di ottenere a unità monetarie e q possibilità di ottenere b unità monetarie, e tutte le possibilità sono ugualmente probabili, allora la mia aspettativa è $(pa + qb)/(p + q)$."

In notazione moderna, se un evento aleatorio può produrre un guadagno x_1 con probabilità p_1 e un guadagno x_2 con probabilità p_2 , allora il valore atteso del guadagno è:

$$E[X] = p_1x_1 + p_2x_2$$

Più generalmente, per una variabile aleatoria discreta che può assumere i valori $x_1, x_2, \dots, x_n$ con le probabilità rispettive $p_1, p_2, \dots, p_n$, il valore atteso è definito come:

$$E[X] = \sum_{i=1 \text{ fino a } n} p_i x_i$$

dove necessariamente $\sum_{i=1 \text{ fino a } n} p_i = 1$, essendo le p_i probabilità che esauriscono tutti i casi possibili.

Questo concetto è fondamentale perché fornisce una misura numerica unica per confrontare diverse opportunità aleatorie. Se devo scegliere tra due giochi d'azzardo diversi, posso calcolare il valore atteso di ciascuno e scegliere quello con il valore atteso maggiore. Ovviamente questa è una semplificazione, come si capirà meglio quando discuteremo il paradosso di San Pietroburgo, ma rappresenta comunque un primo passo cruciale verso una teoria razionale della decisione in condizioni di incertezza.

I Problemi di Huygens e le Equazioni Ricorsive

Il trattato di Huygens è strutturato secondo lo stile classico euclideo, come una serie ordinata di proposizioni accompagnate da dimostrazioni rigorose. Egli presentò quattordici problemi di difficoltà crescente, alcuni dei quali estremamente sofisticati e che richiedono tecniche matematiche avanzate per l'epoca.

Un problema particolarmente interessante e istruttivo è il seguente: supponiamo che tre giocatori A, B e C giochino con due dadi. Il giocatore A vince se ottiene un totale di sei prima degli altri due, il giocatore B vince se ottiene un totale di sette prima degli altri due, e il giocatore C vince se ottiene un totale di otto prima degli altri due. Come dovrebbero essere calcolate le loro aspettative, ossia quanto ciascuno dovrebbe essere disposto a pagare per entrare nel gioco se la posta totale è di 1 unità monetaria?

Per risolvere questo problema, Huygens utilizzò un ragionamento ricorsivo di notevole eleganza. In ogni lancio dei due dadi, le probabilità di ottenere i totali 6, 7 o 8 sono rispettivamente:

$$P(6) = 5/36 \text{ (le coppie } 1+5, 2+4, 3+3, 4+2, 5+1)$$

$$P(7) = 6/36 = 1/6 \text{ (le coppie } 1+6, 2+5, 3+4, 4+3, 5+2, 6+1)$$

$$P(8) = 5/36 \text{ (le coppie } 2+6, 3+5, 4+4, 5+3, 6+2)$$

La probabilità di non ottenere nessuno di questi tre risultati critici è:

$$P(\text{altro}) = 1 - 5/36 - 6/36 - 5/36 = 20/36 = 5/9$$

Se denotiamo con a , b , c le aspettative rispettive dei giocatori A, B, C (supponendo che la posta totale sia 1 unità monetaria), possiamo scrivere le seguenti equazioni ricorsive che esprimono il fatto che l'aspettativa di un giocatore è data dalla somma pesata dei possibili esiti del prossimo lancio:

$$a = (5/36) \times 1 + (5/9) \times a$$

$$b = (6/36) \times 1 + (5/9) \times b$$

$$c = (5/36) \times 1 + (5/9) \times c$$

La prima equazione esprime il fatto che il giocatore A vince immediatamente l'intera posta con probabilità $5/36$ (ottenendo un sei), oppure con probabilità $5/9$ non succede nulla di rilevante e la sua aspettativa rimane invariata al valore a . Le altre equazioni hanno un'interpretazione analoga.

Risolviendo la prima equazione per a :

$$a = 5/36 + (5/9)a$$

$$a - (5/9)a = 5/36$$

$$(4/9)a = 5/36$$

$$a = (5/36) \times (9/4) = 45/144 = 5/16$$

Analogamente per b :

$$b = 6/36 + (5/9)b$$

$$b = 1/6 + (5/9)b$$

$$(4/9)b = 1/6$$

$$b = (1/6) \times (9/4) = 9/24 = 3/8$$

E per simmetria con il giocatore A (stessa probabilità di vincita in ogni lancio), $c = 5/16$.

Possiamo verificare la correttezza del risultato osservando che le tre aspettative devono sommare a 1 (la posta totale deve essere distribuita completamente):

$$a + b + c = 5/16 + 3/8 + 5/16 = 5/16 + 6/16 + 5/16 = 16/16 = 1 \checkmark$$

Questo tipo di analisi, che utilizza le equazioni ricorsive e considera le aspettative come incognite da determinare, fu estremamente innovativo per l'epoca e prefigurò molti sviluppi futuri nella teoria delle decisioni, nella teoria dei giochi, e nello studio dei processi stocastici.

Capitolo IV: Jakob Bernoulli e la Legge dei Grandi Numeri

Dopo i lavori pionieristici di Pascal, Fermat e Huygens nella seconda metà del XVII secolo, il calcolo delle probabilità continuò a svilupparsi principalmente attraverso lo studio di problemi di gioco sempre più complessi e sofisticati. Tuttavia mancava ancora una comprensione profonda e teoricamente fondata di come i risultati teorici del calcolo delle probabilità si collegassero effettivamente alle frequenze osservate nella realtà empirica. Questa lacuna era particolarmente problematica per le applicazioni pratiche: come possiamo usare la teoria per fare inferenze sul mondo reale?

Jakob Bernoulli (1654-1705), matematico svizzero della celebre famiglia Bernoulli che produsse generazioni di matematici eccellenti, dedicò gli ultimi decenni della sua vita a una trattazione sistematica e rigorosa del calcolo delle probabilità. La sua opera magna, *Ars Conjectandi* (Arte di Congetturare), fu pubblicata postuma nel 1713 dal nipote Nikolaus Bernoulli e contiene uno dei risultati più importanti e influenti di tutta la teoria delle probabilità: la legge dei grandi numeri.

La Distribuzione Binomiale

Prima di giungere al suo risultato principale sulla legge dei grandi numeri, Bernoulli dovette sviluppare sistematicamente la teoria delle prove ripetute e delle loro distribuzioni di probabilità. Consideriamo un esperimento che può avere due soli esiti possibili: successo (con probabilità p) o fallimento (con probabilità $q = 1 - p$). Questo tipo di esperimento è quello che oggi chiamiamo una prova di Bernoulli, in onore proprio di Jakob Bernoulli.

Se eseguiamo n prove indipendenti di questo tipo, qual è la probabilità di ottenere esattamente k successi? Bernoulli dimostrò in modo rigoroso che questa probabilità è data dalla formula:

$$P(X = k) = C(n, k) \times p^k \times q^{(n-k)} = [n! / (k!(n-k)!)] \times p^k \times (1-p)^{(n-k)}$$

Questa è la famosa distribuzione binomiale, che oggi denotiamo con la notazione $X \sim \text{Bin}(n, p)$. Il ragionamento alla base di questa formula è il seguente: ci sono $C(n, k)$ modi diversi di scegliere quali k prove tra le n totali saranno successi (e le rimanenti $n-k$ saranno fallimenti). Per ciascuna di queste configurazioni specifiche, la probabilità di ottenere esattamente quella particolare sequenza di successi e fallimenti è $p^k \times q^{n-k}$, poiché le prove sono indipendenti e quindi le probabilità si moltiplicano. Sommando su tutte le $C(n, k)$ configurazioni possibili otteniamo la formula finale.

Il valore atteso del numero totale di successi X è:

$$E[X] = \sum_{k=0}^n k \times C(n, k) \times p^k \times q^{n-k} = np$$

Questo risultato fondamentale può essere dimostrato in vari modi. Il metodo più elegante e concettualmente chiaro è il seguente: se definiamo $X_1, X_2, \dots, X_n$ come le variabili indicatrici (dove $X_i = 1$ se l' i -esima prova è un successo, e $X_i = 0$ altrimenti), allora il numero totale di successi è semplicemente $X = X_1 + X_2 + \dots + X_n$. Per la proprietà di linearità del valore atteso, che vale anche per variabili dipendenti:

$$E[X] = E[X_1] + E[X_2] + \dots + E[X_n] = p + p + \dots + p = np$$

La varianza di X , che misura la dispersione della distribuzione attorno alla media, è:

$$\text{Var}(X) = npq = np(1-p)$$

Questa formula si dimostra usando l'indipendenza delle prove, che implica che la varianza della somma è la somma delle varianze.

Il Teorema di Bernoulli: La Legge dei Grandi Numeri

Il risultato più importante e rivoluzionario dell'Ars Conjectandi è quello che oggi chiamiamo "legge debole dei grandi numeri". Bernoulli dimostrò che se eseguiamo un numero molto grande di prove indipendenti di Bernoulli, allora la frequenza relativa dei successi converge in probabilità alla probabilità teorica di successo. Questo risultato fornisce il ponte cruciale tra la teoria astratta delle probabilità e l'osservazione empirica delle frequenze.

Formalmente, sia X il numero totale di successi in n prove di Bernoulli con probabilità di successo p . Allora, per ogni $\varepsilon > 0$ (arbitrariamente piccolo quanto si vuole):

$$\lim_{n \rightarrow \infty} P(|X/n - p| < \varepsilon) = 1$$

In altre parole, la probabilità che la frequenza relativa X/n differisca dalla probabilità teorica p per più di ϵ tende a zero quando il numero di prove n cresce indefinitamente. Possiamo essere sempre più certi che la frequenza osservata sarà vicina alla probabilità teorica se aumentiamo sufficientemente il numero di osservazioni.

La dimostrazione di Bernoulli, sebbene non rigorosa secondo gli standard moderni di analisi matematica, conteneva sostanzialmente l'essenza dell'argomento corretto. Egli stimò la probabilità che X/n cada in un intervallo centrato in p usando la distribuzione binomiale e mostrò che questa probabilità può essere resa arbitrariamente vicina a 1 scegliendo n sufficientemente grande.

Più precisamente, possiamo riscrivere l'evento di interesse come:

$$P(|X/n - p| \geq \epsilon) = P(|X - np| \geq n\epsilon)$$

Usando la disuguaglianza di Chebyshev, che non era ancora disponibile ai tempi di Bernoulli ma che rende la dimostrazione molto più semplice e diretta, possiamo scrivere:

$$P(|X - np| \geq n\epsilon) \leq \text{Var}(X)/(n\epsilon)^2 = np(1-p)/(n\epsilon)^2 = p(1-p)/(n\epsilon^2)$$

Poiché il prodotto $p(1-p)$ assume il suo valore massimo quando $p = 1/2$, e in quel caso vale $1/4$, abbiamo per ogni $p \in [0,1]$:

$$P(|X/n - p| \geq \epsilon) \leq 1/(4n\epsilon^2)$$

Questa quantità tende manifestamente a zero quando $n \rightarrow \infty$, dimostrando il teorema.

Il Significato Epistemologico del Risultato

Il teorema di Bernoulli ebbe un'importanza che andava ben oltre il suo contenuto matematico specifico, avendo profonde implicazioni epistemologiche e filosofiche. Esso forniva per la prima volta una giustificazione teorica rigorosa per l'uso del metodo empirico nelle scienze: osservando un fenomeno aleatorio un numero sufficientemente grande di volte, possiamo inferire con crescente affidabilità la sua probabilità "vera" o teorica. Questo collegamento fondamentale tra la probabilità teorica (una quantità matematica astratta) e la frequenza osservata (una quantità empiricamente accessibile) fu cruciale per l'applicazione del calcolo delle probabilità alle scienze naturali, alle scienze sociali, e in generale a qualunque dominio in cui si debba ragionare sull'incertezza basandosi su dati empirici.

Bernoulli stesso era profondamente consapevole della portata filosofica del suo risultato. Nell'*Ars Conjectandi* scrisse con eloquenza: "Perfino l'uomo più stupido sa, per un certo istinto naturale e senza alcuna istruzione preventiva (il che è notevole), che più osservazioni si fanno di un fenomeno, meno è il pericolo di errare nel giudicare la sua natura." Il teorema di Bernoulli trasformava questa intuizione vaga e qualitativa in un risultato matematico preciso e quantitativo, mostrando esattamente quanto rapidamente la certezza aumenta con il numero di osservazioni.

Capitolo V: Abraham de Moivre e la Scoperta della Curva Normale

Abraham de Moivre (1667-1754) fu un matematico francese ugonotto che, dopo la revoca dell'Editto di Nantes nel 1685 che pose fine alla tolleranza religiosa in Francia, fu costretto a rifugiarsi in Inghilterra per sfuggire alle persecuzioni religiose. Lì visse una vita economicamente precaria, guadagnandosi da vivere principalmente come consulente matematico per giocatori, speculatori e uomini d'affari, risolvendo problemi complessi di probabilità, annuità e calcoli attuariali.

Nel 1718 pubblicò *The Doctrine of Chances* (La Dottrina delle Probabilità), un'opera monumentale che ampliava e perfezionava sistematicamente i risultati di Huygens e Bernoulli. Ma il suo contributo più significativo e duraturo venne nella seconda edizione del 1738 e in un supplemento del 1733 intitolato *Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi*.

Il Problema delle Approssimazioni

Un problema pratico fondamentale con la distribuzione binomiale è che per valori grandi di n i calcoli diventano estremamente onerosi e praticamente impossibili da eseguire con i mezzi di calcolo disponibili all'epoca. Calcolare il coefficiente binomiale $C(n,k) = n!/(k!(n-k)!)$ per $n = 1000$ richiedeva operazioni con numeri astronomicamente grandi e era sostanzialmente impossibile nell'epoca di de Moivre, quando tutti i calcoli dovevano essere eseguiti manualmente.

De Moivre cercò quindi un'approssimazione utilizzabile per n grande che permettesse di stimare le probabilità binomiali senza dover eseguire calcoli impossibili. Consideriamo la distribuzione binomiale $X \sim \text{Bin}(n,p)$. Abbiamo visto che $E[X] = np$ e $\text{Var}(X) = np(1-p)$. Consideriamo la variabile standardizzata, ossia la variabile trasformata in modo da avere media 0 e varianza 1:

$$Z = (X - np)/\sqrt{np(1-p)}$$

De Moivre scoprì che per n grande la distribuzione di probabilità di questa variabile Z si avvicina a una particolare curva continua, sempre la stessa indipendentemente dai

valori specifici di n e p . Questa curva universale è quella che oggi chiamiamo la distribuzione normale o gaussiana.

La Scoperta della Curva Normale

Nel caso speciale $p = 1/2$ (come nel lancio di una moneta perfettamente equilibrata), de Moivre dimostrò con calcoli laboriosi che per n grande:

$$P(X = k) \approx (1/\sqrt{\pi n/2}) \times \exp[-2(k-n/2)^2/n]$$

In termini della variabile standardizzata $Z = (X - n/2)/\sqrt{n/4} = 2(X - n/2)/\sqrt{n}$, questo diventa:

$$P(a \leq Z \leq b) \approx (1/\sqrt{2\pi}) \int_a^b \exp(-z^2/2) dz$$

Questa è la distribuzione normale standard, che oggi denotiamo con la notazione $N(0,1)$. La funzione densità di probabilità è:

$$\phi(z) = (1/\sqrt{2\pi}) \times \exp(-z^2/2)$$

De Moivre trovò questa formula notevole usando l'approssimazione di Stirling per i fattoriali, scoperta quasi contemporaneamente da James Stirling:

$$n! \approx \sqrt{2\pi n} \times n^n \times \exp(-n)$$

Applicando questa approssimazione alla formula del coefficiente binomiale $C(n,k)$ e facendo diversi passaggi algebrici sofisticati, che richiedono una certa abilità manipolativa e che qui omettiamo per brevità, de Moivre giunse alla formula della curva normale. Il fatto notevole è che questa curva "a campana" emerge naturalmente come approssimazione della distribuzione binomiale, indipendentemente dal valore specifico di n purché sia grande.

L'Estensione al Caso Generale

Nella seconda edizione del 1738 del suo trattato, de Moivre estese il risultato al caso generale di probabilità $p \neq 1/2$ qualsiasi. Il teorema generale, che rappresenta una versione primitiva del teorema del limite centrale, afferma che per $X \sim \text{Bin}(n,p)$:

$$P[(X - np)/\sqrt{np(1-p)} \leq x] \rightarrow \Phi(x) \text{ quando } n \rightarrow \infty$$

dove $\Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^x \exp(-t^2/2) dt$ è la funzione di distribuzione cumulativa normale standard.

Questo risultato è profondo: ci dice che indipendentemente dal valore della probabilità p di successo nelle singole prove, la distribuzione standardizzata del numero totale di successi si avvicina sempre alla stessa curva normale quando n cresce. De Moivre non dimostrò il risultato con il rigore che sarebbe stato richiesto dagli standard moderni dell'analisi matematica, ma la sua intuizione era sostanzialmente corretta e i suoi calcoli erano accurati.

Le Applicazioni Pratiche

De Moivre utilizzò questi risultati per risolvere numerosi problemi pratici che richiedevano il calcolo di probabilità binomiali con n molto grande. Per esempio, nel calcolo delle annuità vitalizie, che era un problema di grande importanza economica e sociale, bisogna stimare la probabilità che una persona di una certa età sopravviva per un dato numero di anni. Usando i dati di mortalità, raccolti per la prima volta in modo sistematico da John Graunt e Edmund Halley, si può modellare la sopravvivenza anno per anno come una sequenza di prove di Bernoulli, e quindi applicare l'approssimazione normale per calcolare le probabilità rilevanti.

Un altro problema affrontato da de Moivre riguardava il gioco della "Pharaon", estremamente popolare nelle sale da gioco dell'epoca. Il problema era tecnicamente molto complesso e richiedeva il calcolo di probabilità binomiali con n molto grande, dell'ordine di centinaia o migliaia. Usando la sua approssimazione normale, de Moivre poteva fornire le risposte accurate in pochi minuti di lavoro, mentre un calcolo diretto usando la formula binomiale esatta avrebbe richiesto ore o addirittura giorni di lavoro manuale estenuante.

Capitolo VI: Pierre-Simon Laplace - La Sintesi Monumentale

Pierre-Simon Laplace (1749-1827) fu uno dei più grandi scienziati dell'epoca napoleonica, paragonabile per statura intellettuale a Newton. Il suo contributo alla teoria delle probabilità fu così vasto e profondo che per molti decenni dopo di lui la disciplina fu essenzialmente "laplaciana". La sua *Théorie Analytique des Probabilités* (1812), un'opera monumentale di oltre 400 pagine, rimase il testo di riferimento fondamentale per tutto il XIX secolo e consolidò la teoria delle probabilità come una branca matura della matematica.

Il Teorema del Limite Centrale nella sua Generalità

Laplace dimostrò una versione molto più generale del risultato di de Moivre, che oggi chiamiamo il teorema del limite centrale. Non solo considerò le somme di variabili di

Bernoulli, ma le somme di variabili aleatorie arbitrarie con opportune condizioni di regolarità.

Sia $X_1, X_2, \dots, X_n$ una sequenza di variabili aleatorie indipendenti con la stessa distribuzione di probabilità, aventi media μ e varianza σ^2 (entrambe finite). Consideriamo la somma $S_n = X_1 + X_2 + \dots + X_n$. Allora per la linearità del valore atteso:

$$E[S_n] = n\mu$$

e per l'indipendenza delle variabili:

$$\text{Var}(S_n) = n\sigma^2$$

La variabile standardizzata:

$$Z_n = (S_n - n\mu)/(\sigma\sqrt{n}) = (\bar{X} - \mu)/(\sigma/\sqrt{n})$$

dove $\bar{X} = S_n/n$ è la media campionaria, ha per costruzione media 0 e varianza 1. Il teorema del limite centrale di Laplace afferma che:

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x)$$

per ogni x reale, dove $\Phi(x)$ è la funzione di distribuzione normale standard. Cioè, per n sufficientemente grande la distribuzione di Z_n si avvicina arbitrariamente alla normale standard, indipendentemente dalla distribuzione originale delle variabili X_i (purché abbiano varianza finita).

Questo risultato è straordinariamente potente e generale. Spiega perché la distribuzione normale appare così frequentemente in natura: ogni fenomeno che è il risultato della somma di molti piccoli effetti indipendenti tenderà naturalmente a seguire una distribuzione approssimativamente normale. Questo è il motivo per cui gli errori di misura, le altezze umane, i punteggi dei test, e innumerevoli altri fenomeni naturali e sociali seguono distribuzioni approssimativamente normali.

La dimostrazione di Laplace utilizzava le funzioni generatrici dei momenti e richiedeva tecniche di analisi matematica molto sofisticate per l'epoca. Sebbene non rigorosa secondo gli standard moderni (in particolare, mancava una giustificazione rigorosa per lo scambio di limite e integrale), la dimostrazione conteneva tutte le idee essenziali che sarebbero state formalizzate rigorosamente nel XX secolo da matematici come Lindeberg e Lévy.

La Formula di Bayes e l'Inferenza Inversa

Uno dei contributi più influenti e controversi di Laplace riguarda quello che oggi chiamiamo "inferenza bayesiana". Il problema fondamentale può essere formulato così: supponiamo di osservare un certo evento o dato E . Questo evento può essere causato o spiegato da diverse ipotesi possibili $H_1, H_2, \dots, H_n$. Come possiamo determinare quale ipotesi è più probabile alla luce dei dati osservati?

Thomas Bayes (1702-1761), un pastore presbiteriano inglese, aveva affrontato questo problema in un manoscritto pubblicato postumo nel 1763 da Richard Price. Tuttavia fu Laplace che sviluppò la teoria in modo sistematico, generale e matematicamente rigoroso, applicandola a innumerevoli problemi pratici di grande importanza.

Il teorema di Bayes nella sua forma più generale afferma:

$$P(H_i | E) = [P(E | H_i) \times P(H_i)] / P(E)$$

dove la probabilità totale dell'evento E è data dalla formula:

$$P(E) = \sum_j P(E | H_j) \times P(H_j)$$

sommando su tutte le ipotesi possibili. In questa formula fondamentale:

- $P(H_i)$ è la probabilità a priori dell'ipotesi H_i , ossia la nostra credenza iniziale prima di osservare i dati E
- $P(E | H_i)$ è la verosimiglianza (likelihood) di osservare E dato che l'ipotesi H_i è vera
- $P(H_i | E)$ è la probabilità a posteriori di H_i , ossia la nostra credenza aggiornata dopo aver osservato E

La dimostrazione del teorema è estremamente semplice e segue direttamente dalla definizione di probabilità condizionata. Dalla definizione:

$$P(H_i | E) = P(H_i \cap E) / P(E)$$

$$P(E | H_i) = P(E \cap H_i) / P(H_i)$$

Poiché per commutatività $P(H_i \cap E) = P(E \cap H_i)$, dalla seconda equazione ricaviamo:

$$P(E \cap H_i) = P(E | H_i) \times P(H_i)$$

Sostituendo nella prima equazione otteniamo immediatamente il teorema di Bayes.

Il Principio della Ragione Insufficiente

Un aspetto filosoficamente controverso dell'approccio di Laplace riguarda la scelta delle probabilità a priori quando non abbiamo informazioni specifiche. Laplace sosteneva il "principio della ragione insufficiente" o "principio di indifferenza": in assenza di informazioni che favoriscano un'ipotesi rispetto a un'altra, tutte le ipotesi devono essere considerate a priori equiprobabili.

Per esempio, nel famoso problema della stima della probabilità di nascita dei maschi, Laplace considerò la probabilità p che nasca un maschio come una variabile incognita da stimare. In assenza di informazioni preliminari, egli assunse una distribuzione uniforme a priori su p , cioè $\pi(p) = 1$ per $p \in [0,1]$, che significa che tutti i valori di p tra 0 e 1 sono considerati inizialmente ugualmente plausibili.

Se poi osserviamo che in n nascite ci sono state k maschi, la verosimiglianza secondo il modello binomiale è:

$$L(p|k,n) = C(n,k) \times p^k \times (1-p)^{(n-k)}$$

Applicando il teorema di Bayes, la distribuzione a posteriori di p è proporzionale al prodotto del prior per la verosimiglianza:

$$\pi(p|k,n) \propto \pi(p) \times L(p|k,n) = 1 \times p^k \times (1-p)^{(n-k)} = p^k \times (1-p)^{(n-k)}$$

Per ottenere una vera densità di probabilità dobbiamo normalizzare dividendo per la costante di integrazione:

$$\pi(p|k,n) = [p^k \times (1-p)^{(n-k)}] / [\int_0^1 t^k \times (1-t)^{(n-k)} dt]$$

L'integrale al denominatore è la funzione Beta di Eulero:

$$B(k+1, n-k+1) = [k!(n-k)!] / [(n+1)!]$$

Quindi la distribuzione a posteriori è:

$$\pi(p|k,n) = [(n+1)! / (k!(n-k)!)] \times p^k \times (1-p)^{(n-k)}$$

Questa è una distribuzione Beta con parametri $\alpha = k+1$ e $\beta = n-k+1$, che scriviamo $Beta(k+1, n-k+1)$. Il valore atteso di questa distribuzione a posteriori è:

$$E[p|k,n] = (k+1)/(n+2)$$

Questo è il famoso stimatore di Laplace o "regola di successione". È interessante notare che differisce leggermente dallo stimatore di massima verosimiglianza $\hat{p} = k/n$,

essendo più "conservativo" nel senso che si sposta sempre verso il valore centrale $1/2$. Questo riflette l'influenza del prior uniforme.

La Regola di Successione di Laplace

Un caso particolare interessante e storicamente famoso si ha quando $k = n$, cioè tutti gli n eventi osservati sono stati successi. Lo stimatore di massima verosimiglianza darebbe semplicemente $\hat{p} = 1$, suggerendo certezza assoluta che l'evento si verificherà sempre. Ma lo stimatore bayesiano di Laplace dà invece:

$$E[p|n,n] = (n+1)/(n+2)$$

che è sempre strettamente minore di 1, per quanto grande sia n . Laplace usò questo risultato per rispondere al problema filosofico: "Il sole è sorto per n giorni consecutivi. Qual è la probabilità che sorga domani?" La sua risposta fu $(n+1)/(n+2)$, che per n grande è molto vicino a 1 ma mai esattamente uguale a 1.

Questo esempio, sebbene apparentemente frivolo e quasi ironico, illustra un punto filosoficamente profondo: anche dopo moltissime osservazioni conformi a un pattern, non possiamo mai essere assolutamente certi matematicamente che il fenomeno continuerà. C'è sempre una piccola probabilità residua di sorpresa, di eccezione, di rottura del pattern osservato. Questa è l'essenza dell'atteggiamento scientifico: apertura permanente alla possibilità di revisione delle nostre credenze alla luce di nuove evidenze.

Le Applicazioni alla Giustizia e alla Scienza

Laplace applicò i suoi metodi probabilistici a problemi di grande rilevanza pratica e sociale. Uno dei più famosi e controversi riguarda l'analisi matematica dei verdetti delle giurie nei tribunali. In Francia all'epoca una giuria di 12 persone decideva sulla colpevolezza o innocenza di un accusato con la regola della maggioranza semplice (7 voti su 12). Laplace analizzò matematicamente la probabilità di errore giudiziario in funzione della "competenza" media dei giurati, definita come la probabilità che un singolo giurato individualmente giudichi correttamente.

Sia p la probabilità che un giurato individualmente giudichi correttamente la colpevolezza o innocenza, e supponiamo che i giurati decidano indipendentemente gli uni dagli altri (un'assunzione chiaramente problematica nella realtà, ma necessaria per l'analisi matematica). La probabilità che almeno 7 su 12 giudichino correttamente è data dalla distribuzione binomiale:

$$P(\text{verdetto corretto}) = \sum_{k=7}^{12} C(12,k) \times p^k \times (1-p)^{(12-k)}$$

Laplace calcolò questa probabilità per diversi valori di p e concluse che se $p > 0,5$ (i giurati sono in media migliori di una scelta casuale), allora la probabilità di un verdetto corretto aumenta significativamente con una giuria di 12 persone rispetto a un singolo giudice. Tuttavia, e questo è il punto critico, se $p < 0,5$ (i giurati sono in media peggiori di una scelta casuale), allora il sistema della giuria può effettivamente peggiorare le cose rispetto a un singolo giudice.

In ambito astronomico, Laplace utilizzò i metodi probabilistici per trattare sistematicamente gli errori di osservazione. Introdusse il metodo dei minimi quadrati (indipendentemente da Gauss, con cui ci fu una famosa disputa di priorità scientifica) per stimare i parametri da dati affetti da errori. Il criterio dei minimi quadrati può essere giustificato probabilisticamente assumendo che gli errori seguano una distribuzione normale: lo stimatore che minimizza la somma dei quadrati degli errori è anche lo stimatore di massima verosimiglianza sotto questa assunzione.

Capitolo VII: Carl Friedrich Gauss e la Teoria degli Errori

Carl Friedrich Gauss (1777-1855), spesso considerato il più grande matematico di tutti i tempi insieme a Archimede e Newton, giunse alla distribuzione normale da una prospettiva completamente diversa rispetto a de Moivre e Laplace. Mentre questi ultimi l'avevano derivata come approssimazione limite della distribuzione binomiale, Gauss la ottenne studiando la teoria degli errori di misura in astronomia e geodesia.

La Teoria degli Errori di Misura

Gauss era profondamente interessato al problema pratico di come combinare ottimalmente le molteplici misurazioni della stessa quantità fisica. Quando misuriamo la posizione di una stella o la lunghezza di un arco di meridiano terrestre dieci volte e otteniamo dieci valori leggermente diversi, qual è la migliore stima della quantità vera? Come dovremmo pesare le diverse misurazioni? Questa domanda aveva enorme importanza pratica per l'astronomia e la geodesia dell'epoca.

Gauss fece le seguenti assunzioni ragionevoli sulla distribuzione degli errori di misura ϵ :

1. Gli errori positivi e negativi della stessa magnitudine sono ugualmente probabili: la densità è simmetrica rispetto a zero
2. Gli errori piccoli sono più probabili degli errori grandi
3. Gli errori estremi sono molto improbabili
4. Gli errori nelle diverse misurazioni sono statisticamente indipendenti

Gauss cercò poi di determinare quale distribuzione di probabilità per gli errori, combinata con il principio della massima verosimiglianza, portasse naturalmente al metodo dei minimi quadrati che egli aveva già sviluppato empiricamente nel 1795, all'età straordinariamente giovane di 18 anni.

Supponiamo che il valore vero della quantità misurata sia θ , e che le misurazioni siano $x_1, x_2, \dots, x_n$. Gli errori sono quindi $\varepsilon_i = x_i - \theta$. Se la funzione densità di probabilità dell'errore è $f(\varepsilon)$, allora la verosimiglianza complessiva delle osservazioni, assumendo indipendenza, è:

$$L(\theta) = \prod_{i=1}^n f(x_i - \theta)$$

Prendendo il logaritmo naturale per semplicità di calcolo:

$$\log L(\theta) = \sum_{i=1}^n \log f(x_i - \theta)$$

Per massimizzare questa verosimiglianza rispetto al parametro θ , deriviamo e uguagliamo a zero:

$$d/d\theta [\log L(\theta)] = -\sum_{i=1}^n [f'(x_i - \theta)/f(x_i - \theta)] = 0$$

Gauss pose la domanda fondamentale: per quale scelta della funzione f questa equazione di massima verosimiglianza è equivalente alla condizione che minimizza la somma dei quadrati degli errori, $\sum_{i=1}^n (x_i - \theta)^2$?

La derivata della somma dei quadrati rispetto a θ è $-2\sum_{i=1}^n (x_i - \theta)$, che si annulla quando θ è uguale alla media aritmetica $\theta = \bar{x} = (1/n)\sum x_i$.

Affinché le due condizioni coincidano per ogni insieme di osservazioni, dobbiamo necessariamente avere:

$$f'(\varepsilon)/f(\varepsilon) = c\varepsilon$$

per qualche costante c da determinare. Questa è un'equazione differenziale ordinaria del primo ordine:

$$f'(\varepsilon) = c \times \varepsilon \times f(\varepsilon)$$

La soluzione generale è:

$$\log f(\varepsilon) = (c/2)\varepsilon^2 + \text{costante}$$

Quindi:

$$f(\varepsilon) = A \times \exp(c\varepsilon^2/2)$$

dove A è una costante di normalizzazione. Per avere una funzione densità di probabilità che decresce per gli errori grandi (in accordo con l'assunzione che gli errori grandi sono improbabili), dobbiamo necessariamente avere $c < 0$. Ponendo $c = -2/\sigma^2$ dove $\sigma > 0$ è un parametro di scala da determinare dai dati:

$$f(\varepsilon) = A \times \exp[-\varepsilon^2/(2\sigma^2)]$$

La costante di normalizzazione A si trova imponendo che l'integrale della densità su tutta la retta reale sia 1:

$$\int_{-\infty}^{\infty} f(\varepsilon) d\varepsilon = A \int_{-\infty}^{\infty} \exp[-\varepsilon^2/(2\sigma^2)] d\varepsilon = 1$$

Usando la formula dell'integrale gaussiano:

$$\int_{-\infty}^{\infty} \exp(-x^2) dx = \sqrt{\pi}$$

con il cambio di variabile appropriato, otteniamo:

$$A = 1/(\sigma\sqrt{2\pi})$$

La densità finale è quindi:

$$f(\varepsilon) = [1/(\sigma\sqrt{2\pi})] \times \exp[-\varepsilon^2/(2\sigma^2)]$$

Questa è la distribuzione normale $N(0, \sigma^2)$. Se $X = \theta + \varepsilon$ è la misurazione (il valore vero più l'errore), allora $X \sim N(\theta, \sigma^2)$ con la densità:

$$f(x) = [1/(\sigma\sqrt{2\pi})] \times \exp[-(x-\theta)^2/(2\sigma^2)]$$

Gauss aveva quindi derivato la distribuzione normale da principi completamente diversi, mostrando che essa emerge naturalmente quando si cerca la distribuzione degli errori che rende il metodo dei minimi quadrati equivalente alla massima verosimiglianza.

Le Proprietà della Distribuzione Normale

La distribuzione normale ha proprietà matematiche notevoli che la rendono centrale nella teoria probabilistica e nelle sue applicazioni:

1. Stabilità per somme: Se $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ sono indipendenti, allora la loro somma è ancora normale: $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Più generalmente, se $X_1, \dots, X_n$ sono variabili normali indipendenti con $X_i \sim N(\mu_i, \sigma_i^2)$, allora ogni combinazione lineare $\sum a_i \times X_i \sim N(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2)$.
2. Le trasformazioni affini: Se $X \sim N(\mu, \sigma^2)$ e $Y = a + bX$ è una trasformazione lineare, allora $Y \sim N(a + b\mu, b^2\sigma^2)$.
3. La standardizzazione: Se $X \sim N(\mu, \sigma^2)$, allora la variabile standardizzata $Z = (X - \mu)/\sigma \sim N(0,1)$ segue la normale standard.
4. La massima entropia: Tra tutte le distribuzioni di probabilità con media μ e varianza σ^2 fissate, la normale è quella con la maggiore entropia informazionale, quindi in un certo senso la più "diffusa" o "meno informativa" compatibilmente con i vincoli dati.
5. La rilevanza del teorema del limite centrale: La normale appare come distribuzione limite delle somme di variabili aleatorie indipendenti, spiegando la sua ubiquità in natura.

Il Lavoro di Gauss sulle Orbite Celesti

La scoperta del primo asteroide Cerere da parte dell'astronomo italiano Giuseppe Piazzi il 1° gennaio 1801 fornì a Gauss l'occasione per dimostrare drammaticamente l'efficacia dei suoi metodi probabilistici. Cerere fu osservata per soli 41 giorni prima di perdersi nel bagliore del sole durante la congiunzione. Con così pochi dati osservazionali, sparsi su un periodo così breve, il calcolo della sua orbita ellittica era estremamente difficile e richiedeva metodi matematici molto sofisticati.

Gauss applicò il metodo dei minimi quadrati combinato con la teoria degli errori normali per calcolare l'orbita completa di Cerere. Le sue previsioni furono straordinariamente accurate: quando Cerere riemerse dalla sua congiunzione solare un anno dopo, nel dicembre 1801, fu ritrovata dall'astronomo Franz Xaver von Zach quasi esattamente nella posizione che Gauss aveva predetto, a meno di mezzo grado di errore angolare.

Questo successo spettacolare portò grande fama internazionale al giovane Gauss, allora solo ventiquattrenne, e consolidò definitivamente il metodo dei minimi quadrati come lo strumento standard per l'analisi statistica dei dati astronomici e geodetici. Il metodo rimane fondamentale ancora oggi in tutte le scienze quantitative.

Capitolo VIII: Siméon-Denis Poisson e i Processi Rari

Siméon-Denis Poisson (1781-1840) fu uno dei più prolifici matematici francesi del XIX secolo, contribuendo a numerosi campi dalla meccanica celeste all'elettromagnetismo. In teoria delle probabilità, estese i risultati dei suoi predecessori in diverse direzioni importanti. Il suo contributo più duraturo e famoso è la distribuzione che porta il suo nome, derivata come caso limite della distribuzione binomiale per eventi rari.

La Distribuzione di Poisson

Consideriamo una situazione in cui abbiamo un numero molto grande di prove n (n grande) ma la probabilità di successo in ciascuna singola prova è molto piccola (p piccolo), in modo tale che il prodotto $np = \lambda$ rimanga costante e finito quando n tende all'infinito. Questa è la situazione tipica degli "eventi rari": molte opportunità per l'evento, ma bassa probabilità in ogni singola opportunità.

Partiamo dalla distribuzione binomiale standard:

$$P(X = k) = C(n, k) \times p^k \times (1-p)^{n-k}$$

Ponendo $p = \lambda/n$ dove λ è una costante positiva:

$$P(X = k) = [n!/(k!(n-k)!)] \times (\lambda/n)^k \times (1-\lambda/n)^{n-k}$$

Possiamo riscrivere questa espressione separando i vari fattori:

$$P(X = k) = (\lambda^k/k!) \times [n(n-1)\dots(n-k+1)/n^k] \times (1-\lambda/n)^n \times (1-\lambda/n)^{-k}$$

Ora esaminiamo il comportamento di ciascun fattore nel limite $n \rightarrow \infty$:

- Il primo fattore $\lambda^k/k!$ è costante rispetto a n
- Il secondo fattore: $n(n-1)\dots(n-k+1)/n^k = [n/n] \times [(n-1)/n] \times \dots \times [(n-k+1)/n] \rightarrow 1$ quando $n \rightarrow \infty$ (prodotto di k termini, ognuno dei quali tende a 1)
- Il terzo fattore: $(1-\lambda/n)^n \rightarrow \exp(-\lambda)$ quando $n \rightarrow \infty$ (per la definizione classica del numero e)
- Il quarto fattore: $(1-\lambda/n)^{-k} \rightarrow 1$ quando $n \rightarrow \infty$

Quindi nel limite otteniamo:

$$\lim_{n \rightarrow \infty} P(X = k) = (\lambda^k/k!) \times \exp(-\lambda) = [e^{-\lambda} \times \lambda^k]/k!$$

Questa è la distribuzione di Poisson con parametro λ , che si denota $X \sim \text{Pois}(\lambda)$.

Le proprietà fondamentali della distribuzione di Poisson sono:

- $E[X] = \lambda$ (il valore atteso è uguale al parametro)
- $\text{Var}(X) = \lambda$ (la varianza è anch'essa uguale al parametro)
- $P(X = k) = [e^{(-\lambda)} \times \lambda^k] / k!$ per $k = 0, 1, 2, \dots$

Una caratteristica notevole è che media e varianza coincidono, il che permette di testare empiricamente se un fenomeno segue effettivamente una distribuzione di Poisson.

Le Applicazioni della Distribuzione di Poisson

La distribuzione di Poisson è appropriata per modellare il numero di eventi rari che si verificano in un intervallo di tempo o spazio fissato, quando questi eventi avvengono indipendentemente e con un tasso medio costante. Gli esempi classici includono:

1. I decadimenti radioattivi: Il numero di particelle emesse da un materiale radioattivo in un intervallo di tempo fissato segue una distribuzione di Poisson. Questo fu uno dei primi fenomeni fisici fondamentali in cui la distribuzione di Poisson fu verificata sperimentalmente.
2. Le chiamate telefoniche: Il numero di chiamate che arrivano a un centralino telefonico in un'ora, quando il numero di potenziali chiamanti è grande ma la probabilità che ciascuno chiami in quel preciso momento è piccola.
3. I difetti di produzione: Il numero di difetti in un rotolo di tessuto di lunghezza fissata, o il numero di errori di stampa in una pagina di un libro.
4. Gli incidenti: Il numero di incidenti stradali in una città in un giorno, o il numero di terremoti in una regione in un anno.

Poisson stesso applicò la distribuzione allo studio statistico dei verdetti giudiziari nelle corti francesi, esaminando il numero di condanne errate, che può essere considerato un evento raro.

Il Processo di Poisson

La distribuzione di Poisson è intimamente connessa con i processi di Poisson, che modellano eventi che si verificano casualmente nel tempo (o nello spazio). Un processo di Poisson con tasso λ (numero medio di eventi per unità di tempo) è caratterizzato dalle seguenti proprietà:

1. Il numero di eventi in intervalli temporali disgiunti sono variabili aleatorie indipendenti
2. Il numero di eventi in un intervallo di lunghezza t segue una distribuzione $\text{Pois}(\lambda t)$
3. La probabilità che due o più eventi si verifichino in un intervallo infinitesimo dt è trascurabile (formalmente, è $o(dt)$ nella notazione di Landau)

Se $N(t)$ denota il numero totale di eventi che si sono verificati fino al tempo t , allora $N(t) \sim \text{Pois}(\lambda t)$, e quindi:

$$P(N(t) = k) = [e^{-(\lambda t)} \times (\lambda t)^k] / k!$$

Un risultato importante è che gli intervalli di tempo tra eventi consecutivi in un processo di Poisson seguono una distribuzione esponenziale con parametro λ . Se T è il tempo di attesa fino al prossimo evento, allora:

$$P(T > t) = e^{-(\lambda t)} \text{ per } t \geq 0$$

e la densità è:

$$f(t) = \lambda e^{-(\lambda t)} \text{ per } t \geq 0$$

La distribuzione esponenziale ha la notevole proprietà di "assenza di memoria": dato che abbiamo già aspettato un tempo s senza che l'evento si verifichi, la distribuzione del tempo di attesa aggiuntivo è ancora esponenziale con lo stesso parametro, indipendentemente da quanto abbiamo già aspettato.

Capitolo IX: L'Emergere della Statistica Matematica

Nella seconda metà del XIX secolo, il calcolo delle probabilità iniziò a trasformarsi gradualmente in statistica matematica. Questa transizione fu guidata dai problemi pratici emergenti in biologia, medicina, scienze sociali ed economia, che richiedevano non solo modelli probabilistici per descrivere fenomeni aleatori, ma anche metodi sistematici per inferire parametri dai dati osservati e testare ipotesi scientifiche. La probabilità, nata come teoria astratta per i giochi d'azzardo, diventava uno strumento concreto per l'indagine scientifica empirica.

Francis Galton e la Correlazione

Francis Galton (1822-1911), cugino di Charles Darwin e poliedrico scienziato vittoriano, fu profondamente affascinato dai fenomeni ereditari e dalla trasmissione dei caratteri da genitori a figli. Studiando sistematicamente le altezze di genitori e figli

in un vasto campione di famiglie inglesi, notò un fenomeno curioso e apparentemente paradossale: i figli di genitori molto alti tendevano ad essere più bassi dei genitori (pur restando comunque sopra la media della popolazione), mentre i figli di genitori molto bassi tendevano ad essere più alti dei genitori (pur restando sotto la media).

Galton chiamò questo fenomeno "regressione verso la mediocrità" (regression toward mediocrity), o più semplicemente "regressione verso la media". Il termine "regressione" è rimasto nella terminologia statistica moderna, anche se oggi non si porta più la connotazione negativa di "mediocrità". Per studiare quantitativamente questo fenomeno, Galton introdusse il concetto fondamentale di coefficiente di correlazione.

Date due variabili aleatorie X e Y (per esempio, l'altezza del padre e l'altezza del figlio), definiamo:

- La media: $\mu_x = E[X]$, $\mu_y = E[Y]$
- La deviazione standard: $\sigma_x = \sqrt{\text{Var}(X)}$, $\sigma_y = \sqrt{\text{Var}(Y)}$
- La covarianza: $\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$

Il coefficiente di correlazione di Pearson (dal nome di Karl Pearson che lo formalizzerà rigorosamente) è definito come:

$$\rho = \text{Cov}(X, Y) / (\sigma_x \times \sigma_y)$$

Questo coefficiente ha le seguenti proprietà fondamentali:

- $-1 \leq \rho \leq 1$ sempre (per la disuguaglianza di Cauchy-Schwarz)
- $\rho = 1$ se e solo se esiste una relazione lineare perfetta positiva: $Y = a + bX$ con $b > 0$
- $\rho = -1$ se e solo se esiste una relazione lineare perfetta negativa: $Y = a + bX$ con $b < 0$
- $\rho = 0$ implica l'assenza di correlazione lineare (ma attenzione: non implica necessariamente l'indipendenza statistica)

Galton stimò empiricamente $\rho \approx 0,5$ tra le altezze di genitori e figli. Questo significava matematicamente che se un padre è 10 cm sopra la media della popolazione, ci si aspetta statisticamente che il figlio sia circa 5 cm sopra la media (la regressione parziale verso la media). Questo fenomeno non implica che la popolazione stia diventando più uniforme nel tempo, ma è semplicemente una conseguenza matematica della natura stocastica dell'ereditarietà.

Karl Pearson e il Chi-quadrato

Karl Pearson (1857-1936), allievo e poi collaboratore di Galton, sviluppò la statistica matematica in modo sistematico e rigoroso, trasformandola in una disciplina scientifica autonoma. Fu il fondatore del primo dipartimento universitario di statistica al mondo, presso l'University College London. Nel 1900 introdusse il test chi-quadrato, uno degli strumenti più importanti e utilizzati dell'inferenza statistica classica.

Il problema fondamentale è il seguente: supponiamo di avere n osservazioni suddivise in k categorie distinte. Osserviamo le frequenze empiriche $O_1, O_2, \dots, O_k$ nelle varie categorie (con $\sum O_i = n$). Una certa teoria scientifica prevede che le frequenze attese siano $E_1, E_2, \dots, E_k$. Come possiamo testare statisticamente se i dati osservati sono compatibili con la teoria, o se invece le discrepanze sono troppo grandi per essere attribuibili al caso?

Pearson propose la statistica:

$$\chi^2 = \sum_{i=1 \text{ fino a } k} [(O_i - E_i)^2 / E_i]$$

Questa statistica misura la discrepanza totale tra frequenze osservate e attese, pesata inversamente per le frequenze attese (in modo che le discrepanze nelle categorie con poche osservazioni attese contino di più). Pearson dimostrò, con un argomento che non era completamente rigoroso secondo gli standard moderni ma era sostanzialmente corretto, che per n grande questa statistica segue approssimativamente una distribuzione chi-quadrato con $k-1$ gradi di libertà.

La funzione densità della distribuzione $\chi^2(v)$ con v gradi di libertà è:

$$f(x) = [1/(2^{v/2} \times \Gamma(v/2))] \times x^{(v/2-1)} \times \exp(-x/2) \text{ per } x > 0$$

dove Γ è la funzione Gamma di Eulero definita da $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \times e^{-t} dt$.

Le proprietà fondamentali della distribuzione chi-quadrato sono:

- $E[\chi^2(v)] = v$ (il valore atteso è uguale ai gradi di libertà)
- $\text{Var}(\chi^2(v)) = 2v$ (la varianza è doppia dei gradi di libertà)
- Se $Z_1, \dots, Z_n \sim N(0,1)$ sono variabili normali standard indipendenti, allora $Z_1^2 + \dots + Z_n^2 \sim \chi^2(n)$

Il test chi-quadrato di Pearson è ancora oggi uno degli strumenti più utilizzati per testare l'adattamento di un modello teorico ai dati osservati.

William Sealy Gosset (Student) e la Distribuzione t

All'inizio del XX secolo, William Sealy Gosset (1876-1937), che lavorava come statistico per la famosa fabbrica di birra Guinness a Dublino, affrontò un problema pratico di grande importanza industriale: come confrontare la qualità di diversi lotti di orzo quando si hanno a disposizione solo pochi campioni per ciascun lotto?

Con campioni piccoli, l'approssimazione normale non è sufficientemente accurata per fare inferenze affidabili. Gosset, che pubblicava sotto lo pseudonimo "Student" perché la Guinness non permetteva ai dipendenti di pubblicare ricerche (temendo che i concorrenti potessero ottenere informazioni proprietarie), derivò la distribuzione esatta della statistica:

$$t = (\bar{X} - \mu) / (S / \sqrt{n})$$

dove $\bar{X}$ è la media campionaria, μ è la vera media della popolazione (supposta normale), S è la deviazione standard campionaria definita da:

$$S^2 = [1/(n-1)] \times \sum_{i=1}^n (X_i - \bar{X})^2$$

e n è la dimensione del campione.

La funzione densità della distribuzione t di Student con v gradi di libertà è:

$$f(t) = [\Gamma((v+1)/2) / (\sqrt{v\pi} \times \Gamma(v/2))] \times (1 + t^2/v)^{-(v+1)/2}$$

Questa distribuzione ha le seguenti proprietà importanti:

- È simmetrica attorno a zero come la normale standard
- Ha code più pesanti della normale (valori estremi sono più probabili)
- Per v grande converge alla normale standard $N(0,1)$
- Per v piccolo differisce significativamente dalla normale

Il lavoro di Student fu fondamentale per lo sviluppo dei metodi statistici applicabili ai piccoli campioni, essenziali in moltissimi contesti pratici dove è impossibile o troppo costoso raccogliere grandi quantità di dati.

Ronald Fisher e l'Inferenza Statistica Moderna

Ronald Aylmer Fisher (1890-1962) è universalmente considerato il padre della statistica moderna. I suoi contributi furono così vasti, profondi e influenti che è difficile esagerarne l'importanza. Trasformò la statistica da una collezione di tecniche

ad hoc in una disciplina scientifica coerente e rigorosa. Menzioneremo solo alcuni dei suoi contributi più importanti.

La massima verosimiglianza: Fisher formalizzò il metodo della massima verosimiglianza come principio generale e ottimale di stima statistica. Data una famiglia di distribuzioni parametrizzata da θ e le osservazioni $x_1, \dots, x_n$, la funzione di verosimiglianza è:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta)$$

Lo stimatore di massima verosimiglianza è il valore del parametro che massimizza questa funzione:

$$\hat{\theta}_{ml} = \operatorname{argmax} L(\theta)$$

Fisher dimostrò che sotto opportune condizioni di regolarità, per n grande lo stimatore $\hat{\theta}_{ml}$ è approssimativamente normale con media θ (il valore vero) e varianza $1/(n \times I(\theta))$, dove $I(\theta)$ è l'informazione di Fisher:

$$I(\theta) = E[-(d^2/d\theta^2) \log f(X; \theta)]$$

Questo risultato fondamentale fornisce non solo uno stimatore, ma anche una misura della sua precisione.

L'analisi della varianza (ANOVA): Fisher sviluppò metodi sistematici per analizzare esperimenti complessi con molteplici fattori. L'idea chiave è decomporre la variabilità totale osservata nei dati nelle componenti attribuibili ai diversi effetti e fattori.

Se abbiamo k gruppi con n_j osservazioni ciascuno, la variabilità totale può essere scritta come:

$$SS(\text{totale}) = \sum_j \sum_i (X_{ij} - \bar{X})^2$$

dove $\bar{X}$ è la media globale di tutte le osservazioni. Questa si decompone matematicamente in:

$$SS(\text{totale}) = SS(\text{tra gruppi}) + SS(\text{entro gruppi})$$

dove:

$$SS(\text{tra gruppi}) = \sum_j n_j (\bar{X}_j - \bar{X})^2 \text{ (variabilità delle medie dei gruppi)}$$

$$SS(\text{entro gruppi}) = \sum_j \sum_i (X_{ij} - \bar{X}_j)^2 \text{ (variabilità dentro ciascun gruppo)}$$

Il test F di Fisher per testare l'ipotesi che tutti i gruppi abbiano la stessa media confronta il rapporto:

$$F = [SS(\text{tra gruppi})/(k-1)] / [SS(\text{entro gruppi})/(N-k)]$$

dove $N = \sum n_j$ è il numero totale di osservazioni. Sotto l'ipotesi nulla (tutti i gruppi hanno la stessa media e la stessa varianza), questa statistica segue una distribuzione F con $(k-1, N-k)$ gradi di libertà.

Il disegno degli esperimenti: Fisher rivoluzionò la sperimentazione scientifica introducendo i concetti fondamentali di randomizzazione, replicazione e blocco. Il suo libro *The Design of Experiments* (1935) stabilì i principi che sono ancora la base della ricerca sperimentale moderna in tutte le scienze. Fisher comprese che un esperimento ben progettato può estrarre molto più informazione dai dati di un esperimento mal progettato con la stessa quantità di osservazioni.

Capitolo X: Verso la Formalizzazione Assiomatica

Nonostante tutti questi progressi straordinari nella teoria e nelle applicazioni, alla fine del XIX secolo mancava ancora una fondazione matematica rigorosa del calcolo delle probabilità. Esistevano diverse interpretazioni filosofiche della probabilità (frequentista, soggettivista, classica laplaciana), e la struttura matematica sottostante non era chiaramente definita in modo univoco e rigoroso.

Il problema era particolarmente acuto per gli spazi infiniti e continui. Per un dado a sei facce possiamo definire la probabilità di ogni faccia come $1/6$ e la probabilità di un qualsiasi insieme di facce come la somma delle probabilità individuali. Ma come definiamo rigorosamente la probabilità sugli insiemi continui come l'intervallo $[0,1]$ o la retta reale completa?

Se assegnassimo probabilità positiva a ogni singolo punto, la somma (integrale) sarebbe infinita. Se assegnassimo probabilità zero a ogni punto, la somma sarebbe zero. Chiaramente serve un concetto matematico diverso: la probabilità deve essere definita per insiemi di punti, non per singoli punti. Ma quali insiemi? E quali proprietà deve soddisfare questa funzione?

Émile Borel e la Teoria della Misura

Émile Borel (1871-1956), matematico francese e figura politica, fece i primi passi decisivi verso una trattazione rigorosa introducendo la teoria della misura. Nel 1898

definì cosa significa matematicamente che un insieme di numeri reali abbia "misura" (lunghezza, area, volume) zero o positiva.

Un insieme $A \subseteq \mathbb{R}$ ha misura di Lebesgue $m(A)$ che soddisfa le seguenti proprietà:

1. $m([a,b]) = b - a$ per ogni intervallo (la misura di un intervallo è la sua lunghezza)
2. $m(A) \geq 0$ per ogni insieme A misurabile (non-negatività)
3. Se $A_1, A_2, \dots$ sono insiemi disgiunti a coppie, allora $m(\bigcup_n A_n) = \sum_n m(A_n)$ (σ -additività)

Borel dimostrò che esistono insiemi "non misurabili", la cui esistenza richiede l'assioma della scelta, ma che la collezione degli insiemi misurabili (chiamata σ -algebra di Borel) è sufficientemente ricca per tutti gli scopi pratici e teorici.

Henri Lebesgue (1875-1941) estese queste idee sviluppando la teoria dell'integrazione di Lebesgue, che permette di integrare funzioni molto più generali e patologiche rispetto all'integrale di Riemann classico. L'integrale di Lebesgue divenne lo strumento matematico essenziale per una teoria rigorosa della probabilità su spazi continui.

La Ricerca di una Struttura Assiomatica

All'inizio del XX secolo il movimento formalista in matematica, guidato da David Hilbert, cercava di fondare tutte le branche della matematica su basi assiomatiche rigorose e logicamente coerenti. La teoria degli insiemi di Zermelo-Fraenkel, la geometria assiomatica, l'algebra astratta erano tutte state formalizzate rigorosamente. La probabilità necessitava dello stesso trattamento.

Diversi matematici proposero assiomatizzazioni parziali. Richard von Mises propose un approccio basato sulle frequenze limite e sui "collettivi". Bruno de Finetti sviluppò una teoria soggettivista basata sulla coerenza delle scommesse. Sergei Bernstein contribuì con importanti risultati sulla convergenza. Ma fu Andrej Kolmogorov che fornì l'assiomatizzazione definitiva che è ancora lo standard oggi.

Capitolo XI: La Fondazione Assiomatica di Kolmogorov

Nel 1933 Andrej Nikolaevich Kolmogorov (1903-1987), allora appena trentenne, pubblicò *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Fondamenti della Teoria della Probabilità), una breve ma densa monografia di circa 60 pagine che trasformò radicalmente e permanentemente la disciplina. In questo lavoro Kolmogorov mostrò come il calcolo delle probabilità potesse essere fondato rigorosamente sulla teoria della misura di Borel-Lebesgue, fornendo così una base matematica unificante che risolse le ambiguità e le controversie precedenti.

Gli Assiomi di Kolmogorov

La formalizzazione di Kolmogorov si basa su tre concetti primitivi fondamentali:

1. Lo spazio campionario Ω : l'insieme di tutti i possibili esiti di un esperimento aleatorio. Gli elementi $\omega \in \Omega$ sono chiamati esiti elementari o punti campionari. Per esempio, nel lancio di un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$; nel lancio di una moneta infinitamente $\Omega = \{H, T\}^\infty$ (sequenze infinite di teste e croci).

2. La σ -algebra $\mathcal{F}$: una famiglia di sottoinsiemi di Ω (chiamati eventi misurabili) che soddisfa:

- $\Omega \in \mathcal{F}$ (l'evento certo è misurabile)
- Se $A \in \mathcal{F}$, allora $A^c = \Omega \setminus A \in \mathcal{F}$ (chiusura per complementazione)
- Se $A_1, A_2, \dots \in \mathcal{F}$, allora $\bigcup_n A_n \in \mathcal{F}$ (chiusura per unioni numerabili)

3. La misura di probabilità P : una funzione $P: \mathcal{F} \rightarrow [0,1]$ che soddisfa i tre assiomi fondamentali:

Assioma 1 (Non-negatività): $P(A) \geq 0$ per ogni $A \in \mathcal{F}$

Assioma 2 (Normalizzazione): $P(\Omega) = 1$

Assioma 3 (σ -additività): Se $A_1, A_2, \dots \in \mathcal{F}$ sono eventi a coppie disgiunti (cioè $A_i \cap A_j = \emptyset$ per $i \neq j$), allora:

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

La terna $(\Omega, \mathcal{F}, P)$ è chiamata spazio di probabilità ed è la struttura fondamentale su cui si basa tutta la teoria moderna delle probabilità.

Le Conseguenze Immediate degli Assiomi

Dagli assiomi seguono immediatamente per via puramente logico-deduttiva diverse proprietà fondamentali che utilizziamo costantemente:

La probabilità dell'evento impossibile: $P(\emptyset) = 0$

Dimostrazione: Consideriamo la sequenza infinita $A_1 = \emptyset, A_2 = \emptyset, A_3 = \emptyset, \dots$. Questi eventi sono disgiunti e la loro unione è ancora l'insieme vuoto: $\bigcup_{n=1}^{\infty} \emptyset = \emptyset$. Per la σ -additività:

$$P(\emptyset) = P(\bigcup_{n=1}^{\infty} \emptyset) = \sum_{n=1}^{\infty} P(\emptyset)$$

Questa uguaglianza può essere vera solo se $P(\emptyset) = 0$ (una serie infinita di termini tutti uguali a un numero positivo divergerebbe).

L'additività finita: Se $A_1, \dots, A_n$ sono eventi disgiunti a coppie, allora:

$$P(\bigcup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$$

Dimostrazione: Definiamo $A_{n+1} = A_{n+2} = \dots = \emptyset$ e applichiamo la σ -additività. L'unione non cambia e i termini aggiunti contribuiscono zero alla somma.

La probabilità del complemento: $P(A^c) = 1 - P(A)$

Dimostrazione: Gli eventi A e A^c sono disgiunti e la loro unione è lo spazio intero: $A \cup A^c = \Omega$. Quindi per additività:

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

Da cui segue immediatamente $P(A^c) = 1 - P(A)$.

La monotonia: Se $A \subseteq B$, allora $P(A) \leq P(B)$

Dimostrazione: Possiamo scrivere B come unione disgiunta $B = A \cup (B \setminus A)$. Quindi:

$$P(B) = P(A) + P(B \setminus A) \geq P(A)$$

poiché $P(B \setminus A) \geq 0$ per l'Assioma 1.

La formula di inclusione-esclusione: Per due eventi qualsiasi A e B :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Dimostrazione: Possiamo scrivere $A \cup B$ come unione disgiunta:

$$A \cup B = A \cup (B \setminus A)$$

Quindi $P(A \cup B) = P(A) + P(B \setminus A)$. Ma possiamo anche scrivere B come unione disgiunta $B = (A \cap B) \cup (B \setminus A)$, quindi:

$$P(B) = P(A \cap B) + P(B \setminus A)$$

Da cui $P(B \setminus A) = P(B) - P(A \cap B)$. Sostituendo otteniamo:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Questa formula si generalizza a n eventi arbitrari nella formula di inclusione-esclusione generale:

$$P(\cup_{i=1}^n A_i) = \sum_i P(A_i) - \sum_{i < j} P(A_i \cap A_j) + \sum_{i < j < k} P(A_i \cap A_j \cap A_k) - \dots + (-1)^{n+1} P(A_1 \cap \dots \cap A_n)$$

La continuità dal basso: Se $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$ è una sequenza crescente di eventi, allora:

$$P(\cup_{n=1}^{\infty} A_n) = \lim_{n \rightarrow \infty} P(A_n)$$

Dimostrazione: Definiamo $B_1 = A_1$, e per $n \geq 2$, $B_n = A_n \setminus A_{n-1}$ (la parte "nuova" di A_n). Allora i B_n sono disgiunti a coppie e $\cup_n B_n = \cup_n A_n$. Inoltre $\cup_{k=1}^n B_k = A_n$. Per la σ -additività:

$$P(\cup_{n=1}^{\infty} A_n) = P(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} P(B_n) = \lim_{n \rightarrow \infty} \sum_{k=1}^n P(B_k) = \lim_{n \rightarrow \infty} P(A_n)$$

La continuità dall'alto: Se $A_1 \supseteq A_2 \supseteq A_3 \supseteq \dots$ è una sequenza decrescente di eventi, allora:

$$P(\cap_{n=1}^{\infty} A_n) = \lim_{n \rightarrow \infty} P(A_n)$$

Dimostrazione: Applicare la continuità dal basso alla sequenza crescente di complementi A_n^c .

Queste proprietà di continuità sono essenziali e distinguono la σ -additività dalla semplice additività finita.

La Probabilità Condizionata e l'Indipendenza

La probabilità condizionata formalizza rigorosamente il concetto intuitivo di "aggiornamento" delle probabilità alla luce di nuova informazione.

Definizione: Se $P(B) > 0$, allora la probabilità condizionata di A dato B è:

$$P(A|B) = P(A \cap B)/P(B)$$

Questa definizione cattura matematicamente l'idea che quando sappiamo che l'evento B si è verificato, lo spazio campionario effettivo si restringe a B , e dobbiamo

rinormalizzare le probabilità di conseguenza. La funzione $P(\cdot|B)$ è essa stessa una misura di probabilità sullo spazio $(B, \mathcal{F} \cap B, P(\cdot|B))$: soddisfa tutti e tre gli assiomi di Kolmogorov con Ω sostituito da B .

La formula delle probabilità totali: Se $B_1, B_2, \dots, B_n$ formano una partizione di Ω (cioè sono disgiunti a coppie e $\cup_i B_i = \Omega$) con $P(B_i) > 0$ per ogni i , allora per ogni evento A :

$$P(A) = \sum_{i=1}^n P(A|B_i) \times P(B_i)$$

Dimostrazione: Poiché i B_i formano una partizione, possiamo scrivere $A = A \cap \Omega = A \cap (\cup_i B_i) = \cup_i (A \cap B_i)$, e gli eventi $A \cap B_i$ sono disgiunti. Quindi:

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i) \times P(B_i)$$

Il teorema di Bayes: Se $P(B) > 0$ e $P(A) > 0$, allora:

$$P(A|B) = [P(B|A) \times P(A)]/P(B)$$

Questo segue immediatamente dalla simmetria di $P(A \cap B) = P(B \cap A)$ e dalla definizione di probabilità condizionata.

L'indipendenza: Due eventi A e B sono indipendenti se:

$$P(A \cap B) = P(A) \times P(B)$$

Se $P(B) > 0$, questa condizione è equivalente a $P(A|B) = P(A)$: sapere che B si è verificato non cambia la probabilità di A . L'indipendenza è quindi la formalizzazione matematica dell'idea che gli eventi non si influenzano reciprocamente.

Più eventi $A_1, \dots, A_n$ sono mutualmente indipendenti se per ogni sottoinsieme non vuoto $I \subseteq \{1, \dots, n\}$:

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$$

È importante notare che l'indipendenza a coppie ($P(A_i \cap A_j) = P(A_i)P(A_j)$ per tutti $i \neq j$) non implica l'indipendenza mutua. Serve che la condizione valga per tutti i sottoinsiemi possibili.

Le Variabili Aleatorie: La Definizione Rigorosa

Una delle innovazioni chiave di Kolmogorov fu la definizione matematicamente rigorosa di variabile aleatoria come funzione misurabile, che prima era un concetto vago e intuitivo.

Definizione: Una variabile aleatoria è una funzione $X: \Omega \rightarrow \mathbb{R}$ tale che per ogni insieme boreliano $B \subseteq \mathbb{R}$ (cioè ogni insieme nella σ -algebra di Borel), l'insieme $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ appartiene alla σ -algebra $\mathcal{F}$.

Questa condizione tecnica di " $\mathcal{F}$ -misurabilità" assicura che possiamo calcolare la probabilità di eventi del tipo $\{X \in B\}$, che è essenziale per tutta la teoria. In particolare, per ogni $x \in \mathbb{R}$ l'insieme:

$$\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$$

quindi possiamo definire senza ambiguità $P(X \leq x)$.

La funzione di distribuzione: La funzione di distribuzione (o funzione di ripartizione) di una variabile aleatoria X è definita come:

$$F_x(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\})$$

Le proprietà matematiche di F_x sono:

1. F_x è non decrescente: se $x_1 \leq x_2$, allora $F_x(x_1) \leq F_x(x_2)$
2. F_x è continua a destra: $\lim_{h \rightarrow 0^+} F_x(x+h) = F_x(x)$
3. $\lim_{x \rightarrow -\infty} F_x(x) = 0$
4. $\lim_{x \rightarrow +\infty} F_x(x) = 1$

Teorema fondamentale: Conversamente, ogni funzione $F: \mathbb{R} \rightarrow [0,1]$ con queste quattro proprietà è la funzione di distribuzione di qualche variabile aleatoria definita su uno spazio di probabilità opportuno.

Le variabili aleatorie discrete: X è discreta se assume al più numerabilmente molti valori $x_1, x_2, x_3, \dots$. La funzione di massa di probabilità è definita come:

$$p_x(x_i) = P(X = x_i) = P(\{\omega : X(\omega) = x_i\})$$

con la condizione necessaria $\sum_i p_x(x_i) = 1$.

Le variabili aleatorie continue: X è (assolutamente) continua se esiste una funzione non negativa $f_x: \mathbb{R} \rightarrow [0, \infty)$ (chiamata densità di probabilità) tale che:

$$F_x(x) = \int_{-\infty}^x f_x(t) dt$$

La densità soddisfa le proprietà:

- $f_x(x) \geq 0$ per ogni x
- $\int_{-\infty}^{\infty} f_x(x) dx = 1$

Per un intervallo $[a, b]$ qualsiasi:

$$P(a \leq X \leq b) = F_x(b) - F_x(a) = \int_a^b f_x(x) dx$$

Nei punti dove F_x è derivabile, $f_x(x) = F'_x(x)$.

Osservazione cruciale: Per una variabile aleatoria assolutamente continua $P(X = x) = 0$ per ogni x specifico. Infatti:

$$P(X = x) = \lim_{h \rightarrow 0^+} P(x \leq X \leq x+h) = \lim_{h \rightarrow 0^+} [F_x(x+h) - F_x(x)] = 0$$

per la continuità di F_x . Questo non significa che l'evento $\{X = x\}$ sia impossibile nel senso logico! Significa solo che ha probabilità nulla. La distinzione tra "impossibile" e "probabilità zero" è sottile ma fondamentale nella teoria moderna.

Il Valore Atteso: Dall'Intuizione all'Integrale di Lebesgue

Il valore atteso (o speranza matematica) formalizza rigorosamente il concetto intuitivo di "valore medio" usando la teoria dell'integrazione di Lebesgue.

Per variabili discrete con valori $x_1, x_2, \dots$:

$$E[X] = \sum_i x_i \times P(X = x_i) = \sum_i x_i \times p_x(x_i)$$

quando la serie converge assolutamente, cioè quando $\sum_i |x_i| \times p_x(x_i) < \infty$.

Per variabili continue con densità f_x :

$$E[X] = \int_{-\infty}^{\infty} x \times f_x(x) dx$$

quando l'integrale converge assolutamente: $\int_{-\infty}^{\infty} |x| \times f_x(x) dx < \infty$.

La definizione generale (via l'integrale di Lebesgue): Per una variabile aleatoria X su $(\Omega, \mathcal{F}, P)$:

$$E[X] = \int_{\Omega} X(\omega) dP(\omega)$$

Questa definizione unifica tutti i casi precedenti (discreto, continuo, e anche casi misti) e si estende a variabili aleatorie che non sono né discrete né continue.

Le proprietà fondamentali del valore atteso:

1. La linearità: Per ogni $a, b \in \mathbb{R}$ e variabili X, Y con valore atteso finito:

$$E[aX + bY] = aE[X] + bE[Y]$$

Questa proprietà vale anche se X e Y sono dipendenti! La linearità è una proprietà fondamentale e potentissima che semplifica moltissimi calcoli.

2. La monotonia: Se $X \leq Y$ quasi certamente (cioè $P(X \leq Y) = 1$), allora:

$$E[X] \leq E[Y]$$

3. Il valore atteso delle costanti: Se c è una costante, allora $E[c] = c$.

4. Il prodotto di variabili indipendenti: Se X e Y sono indipendenti:

$$E[XY] = E[X] \times E[Y]$$

Attenzione: il viceversa non vale! $E[XY] = E[X]E[Y]$ non implica necessariamente l'indipendenza.

Il teorema del cambio di variabile (o teorema del trasporto): Se X ha densità f_x e $g: \mathbb{R} \rightarrow \mathbb{R}$ è una funzione misurabile, allora:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \times f_x(x) dx$$

Per il caso discreto:

$$E[g(X)] = \sum_i g(x_i) \times p_x(x_i)$$

Questo teorema è estremamente utile perché permette di calcolare il valore atteso di funzioni di X senza dover prima determinare la distribuzione di $g(X)$.

La Varianza e i Momenti

La varianza: La varianza misura la dispersione di X attorno alla sua media:

$$\text{Var}(X) = E[(X - E[X])^2]$$

La formula alternativa, spesso più comoda per i calcoli:

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

Dimostrazione: Ponendo $\mu = E[X]$,

$$\begin{aligned}\text{Var}(X) &= E[(X - \mu)^2] = E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2\end{aligned}$$

Le proprietà della varianza:

1. $\text{Var}(X) \geq 0$ sempre
2. $\text{Var}(X) = 0$ se e solo se X è costante quasi certamente
3. $\text{Var}(aX + b) = a^2 \text{Var}(X)$ (la traslazione non cambia la varianza, ma il riscalamento la moltiplica per il quadrato del fattore)
4. Se X e Y sono indipendenti:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Più generalmente, per $X_1, \dots, X_n$ indipendenti:

$$\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n)$$

La deviazione standard: $\sigma_x = \sqrt{\text{Var}(X)}$, ha le stesse unità di misura di X ed è quindi spesso più interpretabile.

La covarianza: Per due variabili X e Y :

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Proprietà:

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ (simmetria)
- $\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y)$
- Se X e Y sono indipendenti, allora $\text{Cov}(X, Y) = 0$ (ma il viceversa non vale!)

Il coefficiente di correlazione:

$$\rho(X,Y) = \text{Cov}(X,Y)/(\sigma_x\sigma_y)$$

dove $\sigma_x = \sqrt{\text{Var}(X)}$, $\sigma_y = \sqrt{\text{Var}(Y)}$.

Proprietà: $-1 \leq \rho \leq 1$, con:

- $\rho = 1$ se e solo se $Y = aX + b$ con $a > 0$ quasi certamente
- $\rho = -1$ se e solo se $Y = aX + b$ con $a < 0$ quasi certamente
- $\rho = 0$ implica incorrelazione (ma non necessariamente indipendenza)

I momenti: Il k-esimo momento di X è:

$$\mu_k = E[X^k]$$

Il k-esimo momento centrato è:

$$\mu'_k = E[(X - E[X])^k]$$

Quindi $\mu'_1 = 0$, $\mu'_2 = \text{Var}(X)$.

Le Disuguaglianze Fondamentali

La disuguaglianza di Markov: Se $X \geq 0$ è una variabile aleatoria non negativa e $a > 0$:

$$P(X \geq a) \leq E[X]/a$$

Dimostrazione: Definiamo la variabile indicatrice $I = 1_{\{X \geq a\}}$ che vale 1 se $X \geq a$ e 0 altrimenti. Allora $X \geq aI$ (perché se $X \geq a$ allora $X \geq a \times 1$, altrimenti $X \geq 0 = a \times 0$). Quindi:

$$E[X] \geq E[aI] = aE[I] = aP(X \geq a)$$

Dividendo entrambi i membri per a otteniamo il risultato.

La disuguaglianza di Chebyshev: Per ogni variabile X con varianza finita e ogni $\varepsilon > 0$:

$$P(|X - E[X]| \geq \varepsilon) \leq \text{Var}(X)/\varepsilon^2$$

Dimostrazione: Appliciamo Markov alla variabile non negativa $Y = (X - E[X])^2$ con $a = \varepsilon^2$:

$$P(|X - E[X]| \geq \varepsilon) = P((X - E[X])^2 \geq \varepsilon^2) \leq E[(X - E[X])^2]/\varepsilon^2 = \text{Var}(X)/\varepsilon^2$$

Forma alternativa: Ponendo $\varepsilon = k\sigma$ dove $\sigma = \sqrt{\text{Var}(X)}$:

$$P(|X - E[X]| \geq k\sigma) \leq 1/k^2$$

Questa forma dice che la probabilità di deviare dalla media per più di k deviazioni standard è al massimo $1/k^2$. Per $k = 2$, almeno il 75% della distribuzione sta entro 2 deviazioni standard dalla media; per $k = 3$, almeno l'89%.

La disuguaglianza di Cauchy-Schwarz: Per variabili X, Y con momenti secondi finiti:

$$|E[XY]| \leq \sqrt{E[X^2]} \times \sqrt{E[Y^2]}$$

con uguaglianza se e solo se $Y = cX$ quasi certamente per qualche costante c .

Conseguenza immediata: $|\text{Cov}(X, Y)| \leq \sigma_X \sigma_Y$, quindi $|\rho(X, Y)| \leq 1$.

La Convergenza di Variabili Aleatorie

Kolmogorov chiarificò rigorosamente i diversi modi in cui una sequenza di variabili aleatorie può convergere, distinguendo tra concetti che erano stati precedentemente confusi.

1. La convergenza quasi certa (q.c.): $X_n \rightarrow X$ quasi certamente se:

$$P(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\}) = 1$$

Cioè, per quasi ogni esito ω (eccetto un insieme di probabilità zero), la sequenza numerica $X_n(\omega)$ converge a $X(\omega)$ nel senso classico dell'analisi.

2. La convergenza in probabilità: $X_n \rightarrow X$ in probabilità se per ogni $\varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P(|X_n - X| > \varepsilon) = 0$$

3. La convergenza in L^p (o in media p -esima): $X_n \rightarrow X$ in L^p se:

$$\lim_{n \rightarrow \infty} E[|X_n - X|^p] = 0$$

Il caso $p = 2$ è particolarmente importante e si chiama convergenza in media quadratica.

4. La convergenza in distribuzione (o in legge): $X_n \rightarrow X$ in distribuzione se:

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

per ogni x in cui F_X è continua.

Le relazioni tra i tipi di convergenza:

1. Convergenza q.c. $\Rightarrow$ convergenza in probabilità
2. Convergenza in $L^p \Rightarrow$ convergenza in probabilità
3. Convergenza in probabilità $\Rightarrow$ convergenza in distribuzione
4. Se $X_n \rightarrow c$ (costante) in distribuzione, allora $X_n \rightarrow c$ in probabilità

Nessuna delle frecce inverse è vera in generale. Queste distinzioni sono fondamentali per comprendere i teoremi limite della teoria delle probabilità.

Le Leggi dei Grandi Numeri

La legge debole dei grandi numeri (WLLN): Sia $X_1, X_2, \dots$ una sequenza di variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) con $E[X_i] = \mu$ e $\text{Var}(X_i) = \sigma^2 < \infty$. Allora:

$$\bar{X}_n = (X_1 + \dots + X_n)/n \rightarrow \mu \text{ in probabilità}$$

Dimostrazione: Per linearità del valore atteso, $E[\bar{X}_n] = \mu$. Per indipendenza, $\text{Var}(\bar{X}_n) = \sigma^2/n$. Applicando Chebyshev:

$$P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \sigma^2/(n\varepsilon^2) \rightarrow 0 \text{ quando } n \rightarrow \infty$$

La legge forte dei grandi numeri (SLLN): Sotto le stesse ipotesi:

$$\bar{X}_n \rightarrow \mu \text{ quasi certamente}$$

$$\text{cioè } P(\lim_{n \rightarrow \infty} \bar{X}_n = \mu) = 1$$

La dimostrazione della legge forte è molto più difficile e richiede tecniche sofisticate. Kolmogorov fornì la prima dimostrazione completamente rigorosa nel 1933.

Il Teorema del Limite Centrale

Il teorema del limite centrale di Lindeberg-Lévy: Sia $X_1, X_2, \dots$ una sequenza i.i.d. con $E[X_i] = \mu$ e $\text{Var}(X_i) = \sigma^2$ dove $0 < \sigma^2 < \infty$. Definiamo:

$$Z_n = (\bar{X}_n - \mu)/(\sigma/\sqrt{n}) = (X_1 + \dots + X_n - n\mu)/(\sigma\sqrt{n})$$

Allora $Z_n \rightarrow N(0,1)$ in distribuzione, cioè per ogni $x \in \mathbb{R}$:

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = \Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^x e^{-t^2/2} dt$$

Questo teorema spiega l'ubiquità della distribuzione normale in natura: ogni fenomeno che è la somma di molti piccoli effetti indipendenti tenderà a seguire una distribuzione approssimativamente normale.

I Processi Stocastici

Definizione: Un processo stocastico è una famiglia $\{X(t) : t \in T\}$ di variabili aleatorie definite sullo stesso spazio di probabilità $(\Omega, \mathcal{F}, P)$. Qui T è l'insieme dei tempi (che può essere discreto come $\mathbb{N}$ o continuo come $[0, \infty)$).

Il moto browniano: Il moto browniano (o processo di Wiener) $W = \{W(t) : t \geq 0\}$ è caratterizzato da:

1. $W(0) = 0$ quasi certamente
2. W ha incrementi indipendenti: per ogni $0 \leq t_1 < t_2 < t_3 < t_4$, gli incrementi $W(t_2) - W(t_1)$ e $W(t_4) - W(t_3)$ sono indipendenti
3. $W(t) - W(s) \sim N(0, t-s)$ per $t > s$
4. I cammini $t \mapsto W(t, \omega)$ sono continui con probabilità 1

Il moto browniano è fondamentale in fisica (moto delle particelle), finanza (modelli di prezzo), e in molti altri campi.

Le martingale: Una sequenza $\{X_n, \mathcal{F}_n\}$ è una martingala se:

1. X_n è $\mathcal{F}_n$ -misurabile (l'informazione al tempo n)
2. $E[|X_n|] < \infty$
3. $E[X_{n+1} | \mathcal{F}_n] = X_n$ quasi certamente

Le martingale formalizzano l'idea di "gioco equo": il valore atteso futuro dato il passato è uguale al valore presente. Sono uno strumento fondamentale della teoria moderna.

Capitolo XII: L'Eredità di Kolmogorov e gli Sviluppi Moderni

L'Impatto della Formalizzazione

L'assiomatizzazione di Kolmogorov ebbe un impatto trasformativo e permanente sul calcolo delle probabilità. Prima del 1933 la disciplina era un misto eterogeneo di intuizione, calcolo esplicito, e applicazioni pratiche, senza una chiara struttura matematica unificante. C'erano ambiguità concettuali, controversie filosofiche sull'interpretazione della probabilità, e mancanza di rigore nei fondamenti. Dopo Kolmogorov, la teoria delle probabilità divenne un ramo pienamente rigoroso della matematica, con teoremi precisi e dimostrazioni formali basate sulla teoria della misura.

Gli aspetti chiave del contributo di Kolmogorov furono:

L'unificazione concettuale: Tutti i diversi approcci alla probabilità (classico laplaciano, frequentista, geometrico) divennero casi speciali della stessa struttura astratta di spazio di probabilità.

La generalità matematica: La formalizzazione permise di trattare spazi infiniti e continui con lo stesso rigore degli spazi finiti, risolvendo problemi che erano stati fonte di confusione.

Il rigore analitico: Le dimostrazioni divennero rigorose secondo gli standard moderni dell'analisi matematica, usando gli strumenti della teoria della misura e dell'integrazione di Lebesgue.

La separazione tra matematica e interpretazione: Kolmogorov separò nettamente la struttura matematica formale dalle questioni interpretative sulla "natura" della probabilità. La matematica è la stessa indipendentemente da come si interpreta filosoficamente $P(A)$.

La connessione con l'analisi moderna: La probabilità divenne strettamente connessa con l'analisi funzionale, la teoria della misura, e la topologia, permettendo l'uso di potenti strumenti matematici moderni.

La base per sviluppi futuri: La formalizzazione fornì la base solida per tutti gli sviluppi successivi: teoria dei processi stocastici, calcolo stocastico, teoria ergodica, probabilità quantistica, e molto altro.

Gli Sviluppi del XX Secolo

Dopo Kolmogorov, la teoria delle probabilità esplose in una miriade di direzioni, diventando uno dei campi più attivi e produttivi della matematica moderna:

Il calcolo stocastico: Kiyosi Itô sviluppò negli anni '40 il calcolo differenziale e integrale per i processi stocastici, con applicazioni fondamentali in fisica e finanza. L'integrale di Itô e il lemma di Itô sono strumenti essenziali nella moderna finanza matematica.

La teoria ergodica: Lo studio dei sistemi dinamici con proprietà statistiche, connettendo probabilità, teoria della misura, e sistemi dinamici.

La teoria delle code: Lo studio matematico dei sistemi di attesa, fondamentale per le telecomunicazioni e la ricerca operativa.

La teoria dell'informazione: Claude Shannon fondò nel 1948 la teoria matematica dell'informazione, usando concetti probabilistici per quantificare l'informazione e la comunicazione.

La probabilità quantistica: Lo sviluppo di versioni non commutative della teoria delle probabilità per la meccanica quantistica.

I metodi computazionali: Gli algoritmi Monte Carlo e i metodi MCMC (Markov Chain Monte Carlo) per approssimazioni numeriche basate sulla simulazione.

Conclusione: Un Viaggio dal Caso al Rigore

Il percorso che abbiamo tracciato, dalle domande di un aristocratico giocatore nel XVII secolo alla formalizzazione assiomatica di Kolmogorov nel XX secolo, illustra magnificamente un pattern ricorrente nella storia della matematica: dai problemi concreti e specifici si sviluppa gradualmente una teoria generale e astratta, che poi si rivela applicabile a domini completamente imprevisi e inimmaginabili al momento della sua nascita.

Il calcolo delle probabilità, nato modestamente per risolvere problemi di giochi d'azzardo e divisione di poste, è diventato il linguaggio fondamentale e insostituibile per descrivere l'incertezza in ogni campo del sapere umano. La formalizzazione di Kolmogorov, lungi dall'essere un esercizio puramente astratto e autoreferenziale, ha permesso lo sviluppo di applicazioni che i pionieri del XVII secolo non avrebbero mai potuto immaginare nei loro sogni più arditi: dalla meccanica quantistica che governa il mondo subatomico alla finanza quantitativa che muove i mercati globali, dall'apprendimento automatico che alimenta l'intelligenza artificiale alla genomica computazionale che decifra il codice della vita.

Eppure, nonostante tutto questo straordinario progresso formale e applicativo, alcune delle questioni filosofiche fondamentali sulla natura ultima della probabilità rimangono aperte, dibattute e affascinanti. Cosa "significa" realmente la probabilità?

È una proprietà oggettiva del mondo fisico, qualcosa che esiste "là fuori" indipendentemente da noi, o è piuttosto una misura soggettiva della nostra ignoranza, uno strumento epistemologico per ragionare in condizioni di incertezza? Come si collega esattamente la probabilità matematica, definita assiomaticamente nello spazio astratto $(\Omega, \mathcal{F}, P)$, alle frequenze concretamente osservate nel mondo reale empirico? Il teorema di Bernoulli fornisce una risposta parziale, ma non risolve completamente il problema filosofico.

La storia del calcolo delle probabilità ci insegna lezioni profonde sulla natura del progresso matematico e scientifico. Ci mostra che la matematica avanza attraverso un continuo e fecondo dialogo tra l'intuizione e il rigore, tra i problemi concreti che emergono dalla pratica e la teoria astratta che li trascende e li unifica, tra le applicazioni pratiche che guidano la ricerca e la riflessione concettuale che chiarisce i fondamenti. È una storia di progressivo e paziente affinamento: dalle nozioni vaghe e intuitive del senso comune alle definizioni precise e inequivocabili, dai calcoli ad hoc e specifici ai teoremi generali e potenti che coprono intere classi di situazioni.

La teoria delle probabilità continua a evolversi, sorprendere, e ispirare. Nuove applicazioni emergono continuamente in campi che si aprono alla frontiera della ricerca: il machine learning moderno si basa pesantemente sui metodi probabilistici bayesiani e sulle reti neurali stocastiche; la crittografia quantistica sfrutta l'intrinseca casualità fondamentale della meccanica quantistica per garantire comunicazioni assolutamente sicure; l'epidemiologia matematica usa processi stocastici sofisticati per modellare la diffusione delle malattie e guidare le politiche di salute pubblica; la teoria dell'informazione algoritmica studia i collegamenti profondi tra casualità, compressione, e complessità.

Dalla sala da gioco del Cavaliere di Méré agli algoritmi che alimentano l'intelligenza artificiale moderna, il calcolo delle probabilità ha percorso una strada lunga, tortuosa, e affascinante. La sua storia è una testimonianza del potere straordinario del pensiero matematico rigoroso, della sua capacità di trasformare intuizioni vaghe in conoscenza precisa, e della sua perenne rilevanza per comprendere il mondo in cui viviamo. In un universo permeato di incertezza, la teoria delle probabilità rimane la nostra bussola più affidabile per navigare attraverso il caso e trovare ordine nel caos.

Appendice

Questa appendice contiene una serie di esempi matematici approfonditi che illustrano i concetti fondamentali della teoria delle probabilità attraverso problemi concreti, dimostrazioni dettagliate e applicazioni pratiche. Ogni esempio è stato scelto per la sua importanza storica o per il suo valore pedagogico nel comprendere aspetti sottili della teoria.

Esempio 1: Il Paradosso di Monty Hall e l'Applicazione di Bayes

Un'applicazione interessante del teorema di Bayes è il famoso paradosso di Monty Hall, che ha causato accesi dibattiti anche tra i matematici professionisti quando fu proposto per la prima volta negli anni '90. Il problema è il seguente:

Sei in un gioco a premi televisivo. Ci sono tre porte: dietro una c'è un'automobile, dietro le altre due ci sono delle capre. Scegli una porta, diciamo la porta 1. Il conduttore del gioco, che sa cosa c'è dietro ogni porta, apre un'altra porta, diciamo la porta 3, rivelando una capra. Ti viene poi data la possibilità di cambiare la tua scelta iniziale. Convieni cambiare?

La risposta intuitiva di molti è che sia indifferente: rimangono due porte, quindi la probabilità è $1/2$ per ciascuna. Questa intuizione è profondamente errata!

Analizziamo con il teorema di Bayes. Sia A_1 l'evento "l'auto è dietro la porta 1", A_2 "l'auto è dietro la porta 2", A_3 "l'auto è dietro la porta 3". Inizialmente:

$$P(A_1) = P(A_2) = P(A_3) = 1/3$$

Sia M_3 l'evento "il conduttore apre la porta 3". Calcoliamo le verosimiglianze:

- $P(M_3 | A_1) = 1/2$ (se l'auto è dietro la porta 1, il conduttore può aprire la porta 2 o 3 con uguale probabilità)
- $P(M_3 | A_2) = 1$ (se l'auto è dietro la porta 2, il conduttore deve aprire la porta 3)
- $P(M_3 | A_3) = 0$ (se l'auto è dietro la porta 3, il conduttore non può aprirla)

La probabilità totale di M_3 è:

$$\begin{aligned} P(M_3) &= P(M_3 | A_1)P(A_1) + P(M_3 | A_2)P(A_2) + P(M_3 | A_3)P(A_3) \\ &= (1/2)(1/3) + (1)(1/3) + (0)(1/3) = 1/6 + 1/3 = 1/2 \end{aligned}$$

Applicando Bayes:

$$P(A_1 | M_3) = [P(M_3 | A_1)P(A_1)]/P(M_3) = [(1/2)(1/3)]/(1/2) = 1/3$$

$$P(A_2 | M_3) = [P(M_3 | A_2)P(A_2)]/P(M_3) = [(1)(1/3)]/(1/2) = 2/3$$

Quindi conviene decisamente cambiare! La probabilità di vincere cambiando è $2/3$, mentre restando sulla scelta iniziale è solo $1/3$.

L'intuizione chiave è che l'azione del conduttore fornisce informazione: aprendo la porta 3, ci sta essenzialmente dicendo "l'auto non è qui, e ho dovuto scegliere tra questa e la porta 2". Se avevamo scelto inizialmente male (probabilità $2/3$), cambiando vinceremo sicuramente.

Esempio 2: Il Problema del Compleanno e le Probabilità Controintuitive

Un classico problema controintuitivo nella teoria delle probabilità: quante persone devono esserci in una stanza affinché la probabilità che almeno due abbiano lo stesso compleanno superi $1/2$?

Assumiamo che i compleanni siano distribuiti uniformemente sui 365 giorni dell'anno (ignorando gli anni bisestili) e che siano indipendenti tra loro.

È più facile calcolare la probabilità complementare: che tutti abbiano compleanni diversi.

Con n persone:

- La prima persona può avere il compleanno in uno qualsiasi dei 365 giorni
- La seconda deve avere un compleanno diverso: probabilità $364/365$
- La terza deve differire dalle prime due: probabilità $363/365$
- ...
- L' n -esima deve differire dalle precedenti $n-1$: probabilità $(365-n+1)/365$

La probabilità che tutti abbiano compleanni diversi è:

$$\begin{aligned} P(\text{tutti diversi}) &= 1 \times (364/365) \times (363/365) \times \dots \times [(365-n+1)/365] \\ &= [365 \times 364 \times \dots \times (365-n+1)]/365^n \\ &= 365!/(365^n \times (365-n)!) \end{aligned}$$

La probabilità che almeno due abbiano lo stesso compleanno è:

$$P(\text{almeno una coincidenza}) = 1 - P(\text{tutti diversi})$$

Calcoliamo per alcuni valori di n :

Per $n = 20$:

$$P(\text{tutti diversi}) \approx 0,5886$$

$$P(\text{coincidenza}) \approx 0,4114$$

Per $n = 23$:

$$P(\text{tutti diversi}) \approx 0,4927$$

$$P(\text{coincidenza}) \approx 0,5073$$

Per $n = 30$:

$$P(\text{tutti diversi}) \approx 0,2937$$

$$P(\text{coincidenza}) \approx 0,7063$$

Sorprendentemente, bastano solo 23 persone per avere una probabilità maggiore di $1/2$ di trovare una coincidenza! Con 30 persone la probabilità sale oltre il 70%.

Questo risultato appare controintuitivo perché non stiamo cercando una persona specifica con lo stesso compleanno di un'altra persona specifica (in quel caso servirebbero molte più persone), ma semplicemente una qualsiasi coppia con lo stesso compleanno tra tutte le $C(n,2) = n(n-1)/2$ possibili coppie. Con 23 persone ci sono 253 coppie diverse da confrontare!

Esempio 3: Il Problema della Rovina del Giocatore

Un giocatore inizia con un capitale a e gioca contro un avversario con un capitale b . A ogni partita il giocatore vince 1 con probabilità p e perde 1 con probabilità $q = 1-p$. Il gioco termina quando uno dei due è in rovina. Qual è la probabilità che il giocatore vinca (cioè arrivi a un capitale totale $a+b$)?

Sia P_i la probabilità di vittoria finale partendo da un capitale i . Chiaramente $P_0 = 0$ (se siamo in rovina, abbiamo perso) e $P_{a+b} = 1$.

Per $0 < i < a+b$, possiamo scrivere un'equazione ricorsiva. Dopo la prima partita ci troviamo nella posizione $i+1$ con probabilità p o nella posizione $i-1$ con probabilità q :

$$P_i = p \times P_{i+1} + q \times P_{i-1}$$

Riarrangiando:

$$p(P_{i+1} - P_i) = q(P_i - P_{i-1})$$

Ponendo $D_i = P_{i+1} - P_i$ (la "differenza prima"), otteniamo:

$$D_i = (q/p)D_{i-1}$$

Quindi:

$$D_i = (q/p)^i D_0$$

Sommando da $i = 0$ a $i = k-1$:

$$P_k - P_0 = D_0 + D_1 + \dots + D_{k-1} = D_0[1 + (q/p) + (q/p)^2 + \dots + (q/p)^{k-1}]$$

Caso 1: Il gioco equo ($p = q = 1/2$)

In questo caso $q/p = 1$, quindi:

$$P_k - P_0 = D_0 \times k$$

$$P_k = k \times D_0$$

Usando $P_0 = 0$ e $P_{a+b} = 1$:

$$1 = (a+b) \times D_0$$

$$D_0 = 1/(a+b)$$

Quindi:

$$P_a = a/(a+b)$$

La probabilità di vittoria è semplicemente proporzionale al capitale iniziale relativo. Se inizio con $a = 30$ e l'avversario ha $b = 70$, la mia probabilità di vincere è $30/100 = 0,30$.

Caso 2: Il gioco non equo ($p \neq 1/2$)

Sia $r = q/p$. La somma geometrica dà:

$$P_k - P_0 = D_0[(1 - r^k)/(1 - r)]$$

Da $P_0 = 0$ e $P_{a+b} = 1$:

$$1 = D_0[(1 - r^{a+b})/(1 - r)]$$

$$D_0 = (1 - r)/(1 - r^{a+b})$$

Quindi:

$$P_a = [(1 - r^a)/(1 - r^{a+b})]$$

Se $r < 1$ (cioè $p > 1/2$, il gioco è favorevole):

- Per a, b finiti, $P_a > a/(a+b)$
- Se $b \rightarrow \infty$, $P_a \rightarrow 1 - r^a$ (la probabilità di non rovinarsi mai)

Se $r > 1$ (cioè $p < 1/2$, il gioco è sfavorevole):

- $P_a < a/(a+b)$
- Se $b \rightarrow \infty$, $P_a \rightarrow 0$ (la rovina è quasi certa)

Questo modello illustra perché i casinò vincono sempre a lungo termine: anche un piccolo vantaggio matematico (p leggermente minore di $1/2$ per il giocatore) porta quasi certamente alla rovina del giocatore se continua a giocare indefinitamente.

Esempio 4: La Stima Bayesiana con il Prior Coniugato

Consideriamo un esempio concreto di inferenza bayesiana. Supponiamo di lanciare una moneta n volte e osservare k teste. Vogliamo stimare la probabilità θ che la moneta dia testa.

L'approccio classico (massima verosimiglianza):

$$\hat{\theta}_{\text{ML}} = k/n$$

L'approccio bayesiano:

Scegliamo un prior Beta(α, β) per θ :

$$\pi(\theta) = [\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))] \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Questo prior ha la media $E[\theta] = \alpha/(\alpha+\beta)$ e rappresenta la nostra credenza iniziale. La scelta di α e β permette di codificare diverse credenze:

- $\alpha = \beta = 1$: il prior uniforme (totale ignoranza)
- $\alpha = \beta = 1/2$: il prior di Jeffreys (invariante per riparametrizzazioni)
- $\alpha = \beta$ grande: una forte credenza che $\theta \approx 1/2$

La verosimiglianza per k teste in n lanci è:

$$L(\theta|k,n) = C(n,k) \times \theta^k(1-\theta)^{n-k}$$

Il posteriore è proporzionale al prodotto prior $\times$ verosimiglianza:

$$\begin{aligned}\pi(\theta|k,n) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \times \theta^k(1-\theta)^{n-k} \\ &= \theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}\end{aligned}$$

Questa è ancora una distribuzione Beta, ora con i parametri $(\alpha+k, \beta+n-k)$. Diciamo che Beta è una famiglia coniugata per la verosimiglianza binomiale.

Il posteriore ha:

- La media: $E[\theta|k,n] = (\alpha+k)/(\alpha+\beta+n)$
- La moda: $(\alpha+k-1)/(\alpha+\beta+n-2)$ per $\alpha, \beta > 1$

Possiamo riscrivere la media posteriore come:

$$E[\theta|k,n] = [(\alpha+\beta)/(\alpha+\beta+n)] \times [\alpha/(\alpha+\beta)] + [n/(\alpha+\beta+n)] \times (k/n)$$

Questa è una media ponderata tra il prior $E[\theta] = \alpha/(\alpha+\beta)$ e la stima MLE k/n . Il peso relativo dipende dalla "forza" del prior $(\alpha+\beta)$ rispetto alla quantità di dati (n) .

Un esempio numerico:

- Il prior: $\alpha = \beta = 2$ (una credenza moderata che $\theta \approx 1/2$)
- I dati: $n = 10$ lanci, $k = 7$ teste
- MLE: $\hat{\theta}_{\{ML\}} = 7/10 = 0,70$
- Il posteriore: Beta(9, 5)
- La media posteriore: $E[\theta|k,n] = 9/14 \approx 0,643$

Con pochi dati il prior ha un'influenza significativa. Se avessimo $n = 100$, $k = 70$:

- La media posteriore: $(2+70)/(2+2+100) = 72/104 \approx 0,692$

Il prior ha ora meno influenza. Nel limite $n \rightarrow \infty$ la stima bayesiana converge alla MLE.

Esempio 5: Il Test di Ipotesi e gli Errori di Tipo I e II

Consideriamo il problema di testare se una moneta è equa basandoci su n lanci. Le ipotesi sono:

- $H_0: \theta = 0,5$ (la moneta equa)
- $H_1: \theta \neq 0,5$ (la moneta non equa)

Sia X il numero di teste osservate. Sotto H_0 , $X \sim \text{Bin}(n, 0,5)$.

Definiamo la regione di rigetto: rifiutiamo H_0 se $|X - n/2|$ è "troppo grande". Più precisamente, fissiamo un livello di significatività α (tipicamente $\alpha = 0,05$) e rifiutiamo H_0 se:

$$|X - n/2| > c$$

dove c è scelto in modo che $P(|X - n/2| > c \mid H_0) = \alpha$.

L'errore di Tipo I: Rifiutare H_0 quando è vera.

$$P(\text{Errore Tipo I}) = P(|X - n/2| > c \mid H_0) = \alpha$$

L'errore di Tipo II: Non rifiutare H_0 quando è falsa.

$$P(\text{Errore Tipo II}) = P(|X - n/2| \leq c \mid H_1) = \beta$$

La potenza del test è $1 - \beta = P(\text{rifiutare } H_0 \mid H_1)$.

Un esempio numerico con $n = 100$:

Sotto H_0 : $E[X] = 50$, $\text{Var}(X) = 25$, $\sigma = 5$

Usando l'approssimazione normale, per $\alpha = 0,05$:

$$Z = (X - 50)/5 \sim N(0,1) \text{ approssimativamente}$$

Per un test a due code rifiutiamo se $|Z| > 1,96$, cioè se:

$$|X - 50| > 1,96 \times 5 = 9,8$$

Quindi rifiutiamo se $X < 40,2$ o $X > 59,8$, cioè (con X intero) se $X \leq 40$ o $X \geq 60$.

Supponiamo ora che la vera probabilità sia $\theta = 0,6$ (H_1 vera). Sotto H_1 :

$$E[X] = 60, \text{Var}(X) = 24, \sigma \approx 4,90$$

La probabilità di non rifiutare (l'errore di Tipo II) è:

$$\begin{aligned} \beta &= P(41 \leq X \leq 59 \mid \theta = 0,6) \\ &\approx P((41 - 60)/4,90 \leq Z \leq (59 - 60)/4,90) \\ &= P(-3,88 \leq Z \leq -0,20) \\ &\approx \Phi(-0,20) - \Phi(-3,88) \\ &\approx 0,4207 - 0,0001 \end{aligned}$$

$$\approx 0,42$$

Quindi la potenza è $1 - \beta \approx 0,58$. Con $n = 100$ e $\theta = 0,6$ abbiamo solo il 58% di probabilità di rilevare che la moneta non è equa.

Questo illustra il trade-off fondamentale nell'inferenza statistica tra gli errori di Tipo I e Tipo II.

Esempio 6: Le Catene di Markov e la Distribuzione Stazionaria

Una catena di Markov discreta è un processo stocastico $\{X_n\}$ che soddisfa la proprietà markoviana:

$$P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i)$$

La catena è caratterizzata dalla matrice di transizione $P = (p_{ij})$ dove $p_{ij} = P(X_{n+1} = j \mid X_n = i)$.

Un esempio: Il modello di clima semplificato

Supponiamo che il tempo possa essere Sole (S) o Pioggia (P), con le transizioni:

- Se oggi c'è sole, domani: sole con probabilità 0,8, pioggia con probabilità 0,2
- Se oggi piove, domani: sole con probabilità 0,4, pioggia con probabilità 0,6

La matrice di transizione è:

$$\begin{array}{cc} & \begin{matrix} S & P \end{matrix} \\ \begin{matrix} S \\ P \end{matrix} & \begin{bmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{bmatrix} \end{array}$$

Se oggi c'è sole, qual è la probabilità che piova tra due giorni?

$$\begin{aligned} P^2(S,P) &= P(S,S) \times P(S,P) + P(S,P) \times P(P,P) \\ &= 0,8 \times 0,2 + 0,2 \times 0,6 \\ &= 0,16 + 0,12 = 0,28 \end{aligned}$$

Calcoliamo P^2 :

$$P^2 = \begin{bmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{bmatrix} \times \begin{bmatrix} 0,8 & 0,2 \\ 0,4 & 0,6 \end{bmatrix} = \begin{bmatrix} 0,72 & 0,28 \\ 0,56 & 0,44 \end{bmatrix}$$

La distribuzione stazionaria:

Una distribuzione $\pi = (\pi_s, \pi_p)$ è stazionaria se $\pi P = \pi$, cioè:

$$\pi_s \times 0,8 + \pi_p \times 0,4 = \pi_s$$

$$\pi_s \times 0,2 + \pi_p \times 0,6 = \pi_p$$

Dalla prima equazione:

$$\pi_s(1 - 0,8) = \pi_p \times 0,4$$

$$0,2\pi_s = 0,4\pi_p$$

$$\pi_s = 2\pi_p$$

Con il vincolo $\pi_s + \pi_p = 1$:

$$2\pi_p + \pi_p = 1$$

$$\pi_p = 1/3$$

$$\pi_s = 2/3$$

A lungo termine, indipendentemente dal tempo di oggi, ci sarà sole 2/3 dei giorni e pioggia 1/3 dei giorni.

Verifichiamo calcolando P^n per n grande:

$$P^4 = \begin{bmatrix} 0,688 & 0,312 \\ 0,624 & 0,376 \end{bmatrix}$$

$$P^8 = \begin{bmatrix} 0,6672 & 0,3328 \\ 0,6656 & 0,3344 \end{bmatrix}$$

$$P^{16} \approx \begin{bmatrix} 0,6667 & 0,3333 \\ 0,6667 & 0,3333 \end{bmatrix}$$

Le righe convergono a $(2/3, 1/3)$, confermando la distribuzione stazionaria.

Il tempo medio di ritorno:

Se siamo in uno stato i , il tempo medio di ritorno a i è $\tau_i = 1/\pi_i$.

Per lo stato S: $\tau_s = 1/(2/3) = 1,5$ giorni

Per lo stato P: $\tau_p = 1/(1/3) = 3$ giorni

Esempio 7: Il Processo di Poisson e i Tempi di Attesa

Consideriamo un centralino telefonico che riceve le chiamate secondo un processo di Poisson con il tasso $\lambda = 3$ chiamate per minuto.

Domanda 1: Qual è la probabilità di ricevere esattamente k chiamate in un intervallo di t minuti?

Sappiamo che $N(t) \sim \text{Pois}(\lambda t)$, quindi:

$$P(N(t) = k) = [e^{-(\lambda t)} \times (\lambda t)^k] / k!$$

Per $t = 5$ minuti e $k = 10$ chiamate:

$$P(N(5) = 10) = [e^{-15} \times 15^{10}] / 10! \\ \approx 0,0486$$

Domanda 2: Qual è il tempo medio di attesa fino alla prossima chiamata?

Il tempo T tra le chiamate consecutive segue una distribuzione esponenziale con il parametro λ :

$$f_T(t) = \lambda e^{-(\lambda t)} \text{ per } t \geq 0$$

Il tempo medio di attesa è $E[T] = 1/\lambda = 1/3$ minuto = 20 secondi.

Domanda 3: Dato che sono passati già 30 secondi senza chiamate, qual è la probabilità di aspettare altri 30 secondi?

Questa è la famosa proprietà di assenza di memoria dell'esponenziale:

$$P(T > s + t \mid T > s) = P(T > t)$$

Il fatto di aver già aspettato non cambia la distribuzione del tempo di attesa futuro!

Esempio 8: La Disuguaglianza di Chernoff e la Concentrazione

Le disuguaglianze di concentrazione sono strumenti potentissimi nella probabilità moderna. Mentre la disuguaglianza di Chebyshev dà:

$$P(|X - E[X]| \geq t) \leq \text{Var}(X)/t^2$$

la disuguaglianza di Chernoff fornisce i limiti esponenzialmente migliori per le somme di variabili indipendenti.

Il teorema (Chernoff): Siano $X_1, \dots, X_n$ le variabili aleatorie indipendenti di Bernoulli con $P(X_i = 1) = p_i$. Sia $X = X_1 + \dots + X_n$ e $\mu = E[X] = \sum p_i$. Allora per ogni $\delta > 0$:

$$P(X \geq (1+\delta)\mu) \leq [e^\delta / (1+\delta)^{(1+\delta)}]^\mu$$

Per $\delta \leq 1$ questo si semplifica in:

$$P(X \geq (1+\delta)\mu) \leq e^{(-\delta^2\mu/3)}$$

La dimostrazione (sketch): Usiamo la tecnica della funzione generatrice dei momenti. Per ogni $s > 0$:

$$P(X \geq (1+\delta)\mu) = P(e^{sX} \geq e^{s(1+\delta)\mu})$$

Per la disuguaglianza di Markov:

$$P(e^{sX} \geq e^{s(1+\delta)\mu}) \leq E[e^{sX}] / e^{s(1+\delta)\mu}$$

Per l'indipendenza:

$$E[e^{sX}] = \prod_{i=1}^n E[e^{sX_i}]$$

Per una Bernoulli con la probabilità p :

$$E[e^{sX}] = (1-p) + pe^s \leq e^{p(e^s-1)}$$

$$\text{Quindi } E[e^{sX}] \leq e^{\mu(e^s-1)}$$

Ottimizzando su s (scegliendo $s = \log(1+\delta)$):

$$P(X \geq (1+\delta)\mu) \leq [e^\delta / (1+\delta)^{(1+\delta)}]^\mu$$

Un esempio numerico: Lanciamo una moneta equa 1000 volte. Qual è la probabilità di ottenere almeno 600 teste?

Qui $n = 1000$, $p = 0,5$, $\mu = 500$, e vogliamo $P(X \geq 600) = P(X \geq 1,2\mu)$, quindi $\delta = 0,2$.

Con Chebyshev: $\text{Var}(X) = 250$, quindi $P(X \geq 600) \leq 250/10000 = 0,025$

Con Chernoff: $P(X \geq 600) \leq e^{(-0,04 \times 500/3)} \approx 0,0013$

Il limite di Chernoff è circa 19 volte migliore!

Esempio 9: Il Problema del Segretario e l'Arresto Ottimale

Questo è un classico problema di decisione sequenziale. Dobbiamo assumere un segretario tra n candidati che si presentano in ordine casuale. Possiamo intervistare i candidati uno alla volta, e dopo ogni intervista dobbiamo decidere immediatamente se assumere (e il processo termina) o rifiutare (e non possiamo tornare indietro). L'obiettivo è massimizzare la probabilità di assumere il migliore.

La strategia ottimale: Rifiutare i primi r candidati (la fase di osservazione), poi assumere il primo candidato successivo che è migliore di tutti quelli visti finora.

Sia $P(r)$ la probabilità di successo con questa strategia. Il candidato i è il migliore e viene assunto se:

1. Il migliore tra i primi $i-1$ è tra i primi r
2. Il candidato i è il migliore complessivo

$$P(r) = (1/n) \times \sum_{i=r+1}^n [r/(i-1)] \\ = (r/n) \times \sum_{j=r}^{n-1} (1/j)$$

Per n grande: $P(r) \approx (r/n) \times \log((n-1)/r)$

Per massimizzare, deriviamo rispetto a r . Ponendo $x = r/n$, otteniamo $x \approx 1/e$.

La strategia ottimale è quindi: rifiutare i primi $n/e \approx 0,368n$ candidati, poi assumere il primo che è migliore di tutti quelli visti.

La probabilità di successo è $P(n/e) \approx 1/e \approx 0,368$

Esempio 10: Le Funzioni Generatrici e il Problema delle Coincidenze

Le funzioni generatrici sono strumenti potenti per risolvere i problemi combinatori probabilistici. Consideriamo il problema delle coincidenze (rencontres): in una permutazione casuale di n oggetti, quanti oggetti si trovano nella loro posizione originale?

Sia X il numero di punti fissi. Per $k = 0$ (derangement):

$$P(X = 0) = \sum_{i=0}^n [(-1)^i / i!] \approx e^{-1} \approx 0,368$$

Per k generale:

$$P(X = k) \approx e^{-1} / k!$$

Questa è la distribuzione di Poisson con il parametro $\lambda = 1$!

Verifica: Usando gli indicatori I_i :

$$X = I_1 + I_2 + \dots + I_n$$

$$E[X] = \sum E[I_i] = n \times (1/n) = 1$$

$$\text{Var}(X) = 1$$

Quindi X ha la media 1 e la varianza 1 per ogni n , esattamente come $\text{Pois}(1)$!